MASTER EN SCIENCES MATHÉMATIQUES

Test de permutations et application au test des Hypervolumes

Blasutig, Laurent

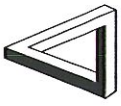
Award date:
2006

FUNDP
NAMUR

Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B - 5000 Napur (Belgique)

Test de permutations et application au test des Hypervolumes



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Promoteur : André Hardy

Laurent BLASUTIG

Année Académique 2005-2006

Je remercie tout d'abord Monsieur Hardy, promoteur de ce mémoire, de m'avoir guidé tout au long de ce travail.

Je remercie également tous mes proches de m'avoir soutenu durant mes études.

Préface

Résumé

Un test d'hypothèse statistique est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.
Pour réaliser un test d'hypothèse, nous avons deux possibilités ; les tests dits classiques et les tests dits de permutations.
Dans la première partie de ce mémoire, nous décrirons ces deux types de méthodes pour ensuite les comparer à travers diverses applications.
Dans la seconde partie, nous présenterons la méthode de classification des Hypervolumes, ainsi que le test des Hypervolumes. Nous utiliserons, pour finir, les tests de permutations de manière à calculer des p-valeurs pour le test des Hypervolumes.

Abstract

A hypothesis test is a statistical procedure to take a decision between two assumptions from results of a sample.
To perform a hypothesis test, we have two possibilities ; tests known as traditional and tests known as of permutations.
In the first part of this memory, we will describe these two methods for then comparing them through various applications.
In the second part, we will introduce the Hypervolumes clustering method and the Hypervolumes test for the number of clusters. To finish we will apply the permutation test in order to compute a p-value for the Hypervolumes test.

Table des matières

Introduction

Les tests d'hypothèse sont la base du raisonnement statistique. Dès le début du $20^{ème}$ siècle, R.A.Fisher proposa la méthode des tests de permutations. Mais ceux-ci demandent beaucoup de calculs ; or à l'époque, on ne disposait pas d'ordinateur. C'est pourquoi, on développa une autre méthode moins fastidieuse pour réaliser un test d'hypothèse : les tests d'hypothèse classiques.

Dans les deux premiers chapitres de ce mémoire, nous présenterons les tests classiques ainsi que les tests de permutations. Nous comparerons, ensuite dans le troisième chapitre, ces deux types de tests à l'aide de quatre applications.

La seconde partie du mémoire traitera d'un des problèmes de la classification qu'est la détermination du nombre de classes. De nombreuses techniques de détermination du nombre de classes ont été proposées et permettent de déterminer le nombre de classes naturelles présentes dans les données sur base de partitions obtenues par des méthodes de classification. Une de ces méthodes est le test des Hypervolumes étant basé sur la méthode de classification des Hypervolumes. Nous présenterons ceux-ci dans, respectivement, le quatrième et cinquième chapitre de cet ouvrage.

L'objectif de la seconde partie de ce mémoire sera de valider deux groupes donnés par une méthode de classification, à l'aide du test des Hypervolumes. Cette validation sera réalisé par un test de permutations.

Pour ce faire, nous implémenterons, dans le sixième chapitre, un programme capable de réaliser un tel test de permutations.

Dans les deux derniers chapitres, nous appliquerons ce test à des données artificielles et présenterons les résultats obtenus par l'algorithme implémenté.

Chapitre 1

Tests classiques

1.1 Introduction

1.1.1 Statistique inférentielle

La statistique inférentielle repose sur une idée simple : il existe un ensemble d'individus appelé population dont les caractéristiques ne sont pas complètement connues. A partir de l'observation d'un sous-ensemble d'individus de cette population, l'échantillon, on va chercher à déterminer, à induire les principales caractéristiques de la population.
La statistique inférentielle élabore des méthodes qui permettent de porter un jugement, de décider à propos de la population, au vu des résultats obtenus pour l'échantillon, en utilisant entre autre le calcul des probabilités.

A partir d'un échantillon donné, on peut calculer différentes valeurs caractéristiques, telles la moyenne ou la variance. De manière plus générale, toute caractéristique, qui peut être calculée à partir des valeurs d'un échantillon aléatoire, est appelée une statistique.
On peut définir, de manière plus formelle, une statistique $T = T(X_1, ..., X_n)$ comme étant une fonction des variables aléatoires observables de l'échantillon $(X_1, ..., X_n)$ qui ne contient aucun paramètre inconnu.

1.1.2 Hypothèse statistique

Rappelons tout d'abord qu'une hypothèse statistique est une affirmation (vraie ou fausse) relative à la population de laquelle est extrait l'échantillon d'observations dont dispose l'expérimentateur.

En se basant sur une théorie, une expérience antérieure ou une argumentation logique, l'expérimentateur est amené, au cours de sa recherche, à formuler une hypothèse au sujet du phénomène qu'il étudie et de la variable qu'il mesure.

Une hypothèse statistique se traduit par un énoncé relatif soit à la forme de la loi de probabilité d'une variable aléatoire, soit de manière plus restrictive à la valeur d'un ou plusieurs paramètres de cette loi. Dans ce dernier cas, on dira que le test est paramétrique.

Une hypothèse est appelée simple quand elle spécifie une seule loi de probabilité. Dans le cas contraire, elle est appelée composite.

1.1.3 Test statistique

Le test d'hypothèse est le processus qui consiste à confronter l'hypothèse émise avec la réalité expérimentale en vue de prendre une décision quant à sa validité (accepter ou rejeter l'hypothèse).

Le rôle d'un test paramétrique statistique est de décider si un certain paramètre d'une population de référence peut prendre ou non une valeur émise par hypothèse, sachant que la statistique correspondante, dont la valeur est calculée à partir d'un échantillon d'objets, peut avoir une valeur quelconque différente.

Dans un test d'hypothèse, la statistique utilisée pour déterminer la région critique est appelée statistique du test. La valeur supposée du paramètre est donnée par l'hypothèse nulle H_0, qui traduit l'hypothèse expérimentale nulle en termes numériques. Le raisonnement se cachant derrière les tests statistiques dérive directement de méthodes scientifiques. En effet, il confronte des résultats expérimentaux, ou des observations, à des constructions théoriques que l'on appelle hypothèses.

1.2 tests d'hypothèse classiques

1.2.1 L'hypothèse nulle et l'hypothèse alternative

Dans les tests d'hypothèse, il est très courant de poser une hypothèse, appelée hypothèse nulle H_0, qui va directement à l'encontre de ce que l'on espère démontrer. H_0 sera, très souvent, une affirmation de "non différence" (par exemple $H_0 : \mu = \mu_0$).

On recourt souvent à l'hypothèse nulle pour plusieurs raisons :

• D'un point de vue théorique d'abord, il est plus facile de démontrer l'inexactitude d'une hypothèse, alors que prouver son exactitude est souvent virtuellement impossible.
Par exemple, admettons que je veuille tester l'hypothèse "Toutes les baleines vivantes ont un estomac ". Pour démontrer l'exactitude de cette hypothèse, il faudrait que j'examine toutes les baleines et que je trouve un estomac dans chacune d'elles. Et encore, cela ne me dirait rien sur les cétacés à naître et déjà disparus. Par contre, il suffirait que je trouve une seule baleine vivante sans estomac pour prouver l'inexactitude de l'hypothèse.

• D'un point de vue pratique, on travaille souvent avec l'hypothèse nulle car elle peut servir de point de départ aux calculs de probabilités. (En effet, l'hypothèse nulle étant une affirmation de "non différence", on peut ainsi, sous H_0, remplacer le paramètre du test par une valeur donnée par H_0.)

On pose souvent une hypothèse nulle accompagnée d'une hypothèse alternative, ainsi, si les résultats expérimentaux sont tels qu'ils nous forcent à rejeter l'hypothèse nulle, nous pourrons alors accepter que son contraire est au moins plausible. C'est l'hypothèse alternative H_1, que nous accepterons si l'hypothèse nulle est rejetée.

Remarquons qu'en statistique, on ne peut pas prouver que l'hypothèse H_1, qui représente le contraire de l'hypothèse principale, est vraie. On ne peut pas accepter l'hypothèse contraire comme conséquence logique du rejet de l'hypothèse principale (sauf lorsque l'hypothèse contraire est très générale). Le test statistique ne teste que H_0 ; on accepte ou rejette H_0.

1.2.2 Test unilatéral et bilatéral

Un autre aspect d'un test statistique, c'est l'hypothèse alternative (H_1), qui est elle aussi imposée par le problème initial. H_1 est l'opposé de H_0, mais il peut y avoir plusieurs manières de représenter l'opposé de H_0.

Prenons pour exemple le test suivant, où l'on s'intéresse à la moyenne d'une population. On veut montrer que cette moyenne est différente de μ_0.
Si l'on veut montrer que celle-ci est différente de μ_0 dans les deux directions (plus grande que μ_0 et plus petite que μ_0) alors une hypothèse alternative bilatérale est constituée à propos de la valeur du paramètre de la statistique de la population : $\mu \neq \mu_0$.
Ou, au contraire, si le problème à l'origine de l'hypothèse impose soit obligatoirement plus grande ou plus petite que μ_0, on formulera alors une hypothèse alternative unilatérale (par exemple $\mu < \mu_0$ unilatéral à gauche).

Le terme de bilatéral vient du fait que lorsque nous comparerons la valeur observée de la statistique du test, à la distribution des valeurs qu'elle aurait pu connaître si H_0 était vraie, nous considérerons les deux extrémités de cette distribution comme des zones de rejet de H_0 .

Par opposition, des tests sont dits unilatéraux parce que la zone de rejet de H_0 se situe à une seule extrémité de la distribution de probabilités qui nous sert de référence.

Dans les tests statistiques paramétriques classiques, la statistique du test calculée à partir des données se réfère à des distributions souvent utilisées comme par exemple les distributions Z, t, F et χ^2. Ceci ne peut cependant avoir lieu que si certaines conditions sont remplies par les données. Les conditions les plus souvent rencontrées sont celles de la normalité de la (des) variable(s) de la population de référence, d'homoscédasticité et d'indépendance des observations.

Remarquons aussi que certains tests particuliers (test F en analyse de variance, test χ^2) sont cependant toujours unilatéraux, de par leur construction particulière.

1.2.3 Erreurs de première et seconde espèce

Soit $X_1, ..., X_n$ un échantillon aléatoire simple d'une distribution $f(x;\theta)$ et γ un test statistique d'une hypothèse H_0 (par exemple $\theta = \theta_0$), défini par "on rejette H_0 ssi $\{x_1, ..., x_n\} \in W_\gamma$" où W_γ est la région critique du test γ.

Si à l'issue de ce test, on rejette une hypothèse H_0 qui n'aurait pas dû être rejetée, on dit que l'on commet une erreur de type I. Dès lors, la probabilité de l'erreur de type I, appelée risque de première espèce, est égale à :

$$\alpha_\gamma(\theta) = P_{H_0}(W_\gamma).$$

Par contre, si on ne rejette pas une hypothèse H_0 qui aurait dû être rejetée (ou encore, rejeter H_1, alors que H_1 est vraie), on commet une erreur de type II. Le risque de deuxième espèce sera donc la probabilité de l'erreur de type II, donnée par :

$$\beta_\gamma(\theta) = P_{H_1}(\overline{W}_\gamma).$$

On peut résumer tout cela grâce au tableau ci-dessous :

	H_0 est vraie	H_1 est vraie
On ne rejette pas H_0	Décision correcte	Erreur de type II
On rejette H_0	Erreur de type I	Décision correcte

Remarque : Le chercheur connaît toujours la valeur de α puisque c'est lui qui l'a déterminée; dans bien des cas cependant, on ne connaît pas précisément β (notion de puissance d'un test) qui doit être déterminé par une analyse de puissance.

1.2.4 Puissance d'un test et niveau de signification

On appelle puissance d'un test la probabilité de rejeter H_0 lorsque H_1 est vraie (rejet avec raison) :

$$\Pi_\gamma(\theta) = P_{H_1}(W_\gamma) = 1 - \beta_\gamma(\theta) .$$

On définira aussi, la courbe d'efficacité d'un test γ, qui nous renseigne sur les performances d'un test :

$$e_\gamma(\theta) = P_\theta(\text{rejeter } H_0) .$$

Supposons que nous ayons un test γ de l'hypothèse $H_0 : \theta \in \Theta_0$, où $\Theta_0 \subset \Theta$ et Θ représente l'ensemble des valeurs possibles de θ. Alors, on définira le niveau du test γ par :

$$\alpha^* = \sup_{\theta \in \Theta_0} e_\gamma(\theta) .$$

De plus, si H_0 est une hypothèse simple, alors le niveau du test γ est aussi appelé le niveau de signification du test γ : $\alpha^* = \sup_{\theta \in \Theta_0} e_\gamma(\theta) = e_\gamma(\theta_0) = \alpha$.

Le niveau de signification α d'un test détermine les $\alpha \cdot 100\%$ des échantillons les plus "défavorables" à H_0. On rejette H_0, au niveau α, si l'échantillon observé appartient à l'ensemble de ces $\alpha \cdot 100\%$ échantillons les plus "incompatibles" avec H_0. Le niveau de signification doit être choisi en fonction de la force des preuves contre l'hypothèse nulle.

En testant une hypothèse, le chercheur doit déterminer quelle est la probabilité d'erreur α qu'il est prêt à tolérer. Ce choix est arbitraire, mais on emploie la plupart du temps les seuils de signification $\alpha = 0{,}05$ (résultat significatif), 0,01 (hautement significatif) ou 0,001 (très hautement significatif) ; les seuils de 5% et 1% furent proposés par Fisher (1925). Si on décide de réaliser un test au seuil $\alpha = 5\%$ par exemple, cela signifie que l'on se donne 5 chances sur 100 de rejeter H_0 même si H_0 est vraie et ne devrait donc pas être rejetée. Ceci montre bien que le seuil α doit être choisi en fonction de la gravité des conséquences que l'on encourra si on est amené, par le test, à prendre la mauvaise décision.

Ainsi, si des sommes d'argent importantes sont en jeu, par exemple pour le lancement d'un nouveau médicament qui serait plus efficace dans le traitement d'une maladie que les médicaments déjà sur le marché, il convient d'employer un seuil de signification α extrêmement petit lors des tests statistiques qui établissent les qualités supérieures du nouveau produit, de façon à réduire le risque de se tromper.

Par contre, lors des tests d'une procédure médicale nouvelle permettant peut-être de traiter une maladie jusque là incurable, on peut se permettre une probabilité d'erreur α relativement grande, tout effet du traitement, même partiel, représentant un progrès. Une augmentation du risque d'erreur α entraîne une réduction du risque d'erreur β.

1.3 Règles de décision

Pour résoudre un test d'hypothèse, nous avons vu lors du cours de *statistiques* en seconde candidature deux méthodes qui permettent de décider si il faut ou non rejeter l'hypothèse H_0 et qui donnent des conclusions identiques.

L'une calcule la région critique du test tandis que l'autre se base sur la probabilité sous H_0 d'avoir collecté un échantillon au moins aussi "rare" que celui obtenu dans les données du test.

1.3.1 Règle de décision avec région critique.

Les deux premières étapes d'un test d'hypothèse (formuler les hypothèses et fixer le niveau de signification α) ont déjà été passées en revue. Il faut maintenant déterminer la statistique du test ainsi que la distribution de celle-ci sous H_0 (par exemple : Z, t, F, χ^2).

Ensuite, on définira la région critique du test W (zone de rejet et zone d'acceptation) grâce par exemple au théorème de Neyman-Pearson ou à celui de Lehmann.

Il ne reste plus dès lors qu'à établir la règle de décision. La décision statistique est prise en comparant la valeur observée de la statistique du test à la distribution des valeurs que l'on pourrait obtenir sous H_0 ; ces valeurs sont fournies par les tables des différentes lois de distribution. On rejettera H_0 au niveau de signification α ssi $(x_1, ..., x_n) \in W$.

Remarquons que pour pouvoir utiliser les lois théoriques, les échantillons doivent bien souvent vérifier certaines hypothèses. En effet, les statistiques du test obéissent à certaines lois lorsque Ho est vraie, sous certaines conditions qui dépendent des tests. Par exemple, dans de nombreux tests, la loi de distribution ne s'applique au cas où H_0 est vraie que si les données sont extraites d'une population (dont la distribution est) Normale. Une autre condition très importante est l'indépendance des observations.

1.3.2 La p-valeur

La p-valeur d'un test statistique est définie comme la probabilité d'observer, sous H_0, un échantillon au moins aussi "extrême" que celui qui a été observé. Plus la p-valeur est petite, plus l'échantillon observé sera "extrême" par rapport à H_0, et plus on sera convaincu de la pertinence du rejet de H_0.

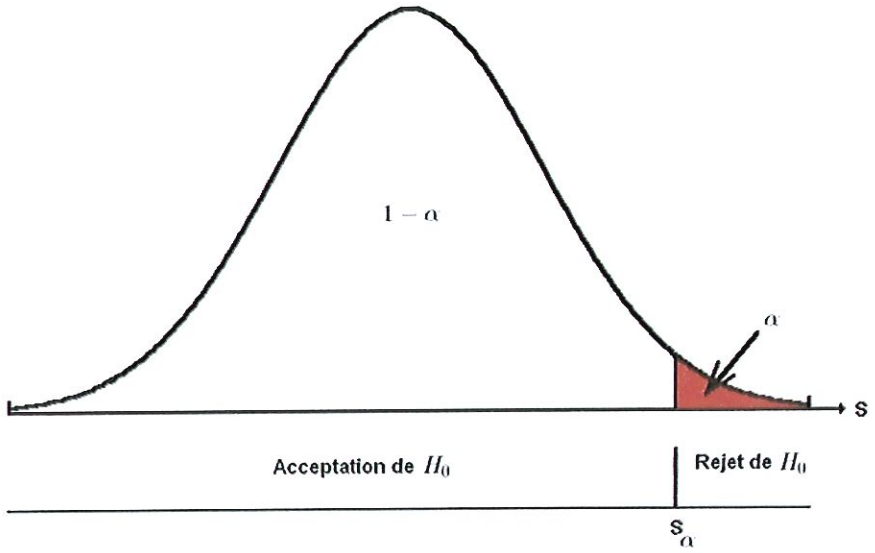
La zone dite "d'acceptation de H_0" est l'intervalle des valeurs de la statistique du test dans lequel les différences observées peuvent être attribuées aux variations dues à l'échantillonnage. La zone de rejet est au contraire la zone dans laquelle la statistique du test prend une valeur trop extrême pour qu'on puisse l'attribuer à une variation aléatoire prévisible sous l'hypothèse nulle (H_0). Dans ce cas, on rejettera H_0 avec, bien sûr, un risque d'erreur maximal égal à la valeur α choisie par l'expérimentateur.

Pour une statistique S quelconque, les zones "d'acceptation" et de rejet ne sont pas les mêmes, selon que le test est unilatéral ou bilatéral. S_α et $S_{\alpha/2}$ sont les valeurs critiques, tirées d'une table de la loi de S pour respectivement les tests unilatéraux et bilatéraux . Dans tous les cas, $(1-\alpha)$ est la probabilité de prendre la bonne décision si H_0 est vraie.

La règle de décision, en fonction du type de test, sera donc la suivante :

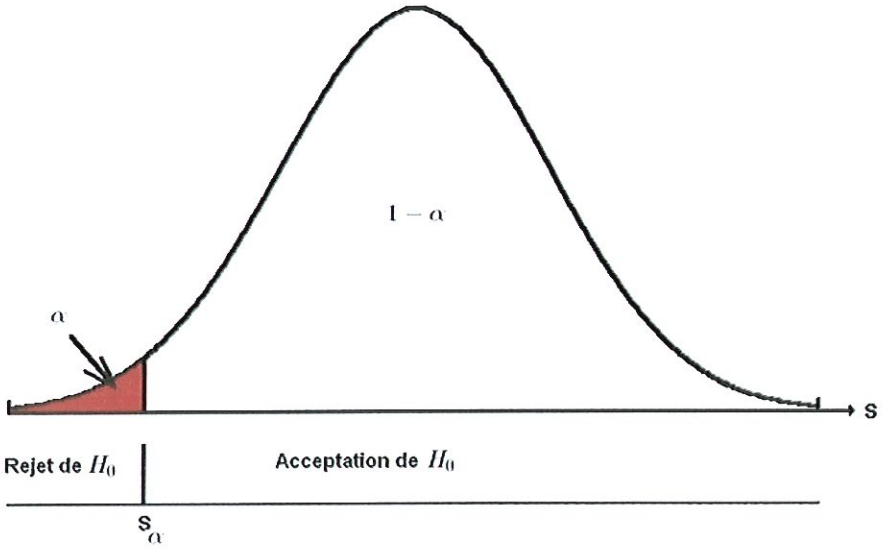
• Pour un test unilatéral à droite, on rejette H_0 au niveau de signification α, si la p-valeur est inférieure à α ou encore si :

$$P_{H_0}(S \geq s_{obs}) \leq \alpha$$



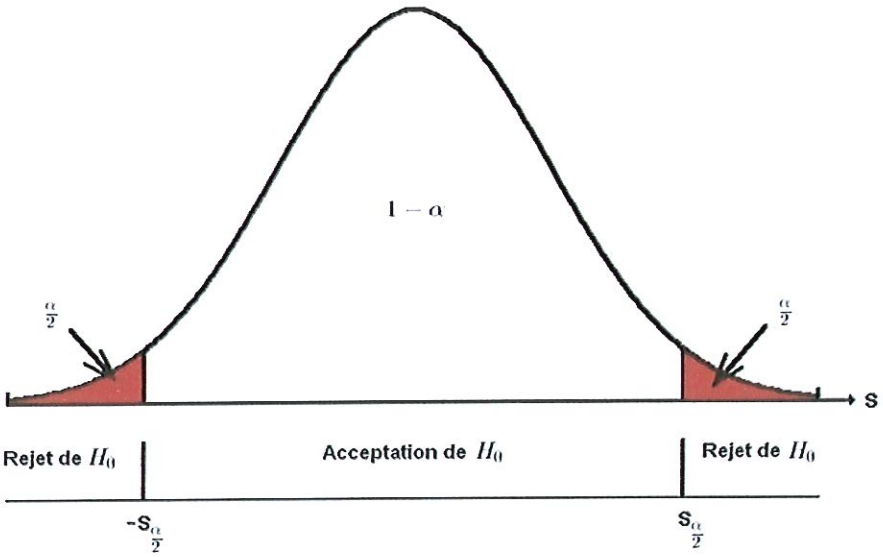
• Pour un test unilatéral à gauche, on rejette H_0 au niveau de signification α, si la p-valeur est inférieure à α ou encore si :

$$P_{H_0}(S \leq s_{obs}) \leq \alpha$$



• Pour un test bilatéral, on rejette H_0 au niveau de signification α, si les p-valeurs sont inférieures à $\frac{\alpha}{2}$ ou encore si :

$$P_{H_0}(S \leq s_{obs}) \leq \frac{\alpha}{2} \quad \text{ou si} \quad P_{H_0}(S \geq s_{obs}) \leq \frac{\alpha}{2}$$



Remarque :

$S_{\frac{\alpha}{2}} = -S_{1-\frac{\alpha}{2}}$ uniquement si la distribution est symétrique.

1.3.3 Test relatif à la différence des moyennes de deux populations Normales (avec σ_1 et σ_2 connus)

On suppose les échantillons aléatoires issus respectivement d'une population $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, de taille n_1 et n_2 où σ_1 et σ_2 sont connus. Les échantillons sont supposés indépendants.

• Considérons les trois problèmes de test suivants :

$$a) \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} \text{(test unilatéral à droite)}$$

$$b) \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases} \text{(test unilatéral à gauche)}$$

$$c) \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \text{(test bilatéral)}$$

• La statistique du test est donnée par :

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sous H_0, $Z \sim N(0,1)$.

• Région critique : on rejette H_0 au niveau de signification α si

a) $z_{obs} \geq k_\alpha$ où k_α est défini par

$$P_{H_0}(Z \geq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, z \geq k_\alpha\}.$$

b) $z_{obs} \leq k_\alpha$ où k_α est défini par

$$P_{H_0}(Z \leq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, z \leq k_\alpha\}\,.$$

c) $z_{obs} \leq k_{1,\alpha}$ ou si $z_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(Z \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(Z \geq k_{2,\alpha}) = \frac{\alpha}{2}.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, z \leq k_{1,\alpha} \text{ ou } z \geq k_{2,\alpha}\}\,.$$

- Calcul de la p-valeur : on rejette H_0, au niveau de signification α, si

  a) $P_{H_0}(Z \geq z_{obs}) \leq \alpha$

  b) $P_{H_0}(Z \leq z_{obs}) \leq \alpha$

  c) $P_{H_0}(Z \leq z_{obs}) \leq \frac{\alpha}{2}$ ou si $P_{H_0}(z \geq z_{obs}) \leq \frac{\alpha}{2}$.

1.3.4 Test relatif à la différence des moyennes de deux populations Normales (avec $\sigma_1 = \sigma_2$ inconnus)

On suppose les échantillons aléatoires issus respectivement d'une population $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, de taille n_1 et n_2 où σ_1 et σ_2 sont inconnus. Les échantillons sont supposés indépendants.

- Considérons les trois problèmes de test suivants :

  a) $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$ (test unilatéral à droite)

  b) $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$ (test unilatéral à gauche)

  c) $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$ (test bilatéral)

• La statistique du test est donnée par :

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}}}$$

où $S_j^2 = \dfrac{\sum_{i=1}^{n_j} \left(X_i - \overline{X}\right)^2}{n_j - 1}$, $j \in \{1, 2\}$.

Sous H_0, $T \sim t_{n_1+n_2-2}$.

• Région critique : on rejette H_0 au niveau de signification α si

a) $t_{obs} \geq k_\alpha$ où k_α est défini par

$$P_{H_0}(T \geq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \geq k_\alpha\}.$$

b) $t_{obs} \leq k_\alpha$ où k_α est défini par

$$P_{H_0}(T \leq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq k_\alpha\}.$$

c) $t_{obs} \leq k_{1,\alpha}$ ou si $t_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(T \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(T \geq k_{2,\alpha}) = \frac{\alpha}{2}.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq k_{1,\alpha} \text{ ou } t \geq k_{2,\alpha}\}.$$

• Calcul de la p-valeur : on rejette H_0, au niveau de signification α, si

a) $P_{H_0}(T \geq t_{obs}) \leq \alpha$

b) $P_{H_0}(T \leq t_{obs}) \leq \alpha$

c) $P_{H_0}(T \leq t_{obs}) \leq \frac{\alpha}{2}$ ou si $P_{H_0}(T \geq t_{obs}) \leq \frac{\alpha}{2}$.

1.3.5 Test relatif à un rapport de variances (μ_1 et μ_2 inconnus)

On suppose les échantillons aléatoires issus respectivement d'une population $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, de taille n_1 et n_2 où μ_1 et μ_2 sont inconnus. Les échantillons sont supposés indépendants.

• Considérons les trois problèmes de test suivants :

$$a) \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{cases} \quad \text{(test unilatéral à droite)}$$

$$b) \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases} \quad \text{(test unilatéral à gauche)}$$

$$c) \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \quad \text{(test bilatéral)}$$

• La statistique du test est donnée par :

$$F = \frac{\frac{(n_1 - 1)}{\sigma_1^2} S_1^2}{\frac{(n_2 - 1)}{\sigma_2^2} S_2^2}$$

où $S_j^2 = \dfrac{1}{n_j - 1} \displaystyle\sum_{i=1}^{n_j} \left(X_i - \overline{X}\right)^2$, $j \in \{1, 2\}$.

Sous H_0, $F \sim F_{n_1 - 1, n_2 - 1}$.

• Région critique : on rejette H_0 au niveau de signification α si

a) $f_{obs} \geq k_\alpha$ où k_α est défini par

$$P_{H_0}(F \geq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, f \geq k_\alpha\}\,.$$

b) $f_{obs} \leq k_\alpha$ où k_α est défini par

$$P_{H_0}(F \leq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, f \leq k_\alpha\}\,.$$

c) $f_{obs} \leq k_{1,\alpha}$ ou si $f_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(F \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(F \geq k_{2,\alpha}) = \frac{\alpha}{2}\,.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, f \leq k_{1,\alpha} \text{ ou } f \geq k_{2,\alpha}\}\,.$$

• Calcul de la p-valeur : on rejette H_0, au niveau de signification α, si

a) $P_{H_0}(F \geq f_{obs}) \leq \alpha$

b) $P_{H_0}(F \leq f_{obs}) \leq \alpha$

c) $P_{H_0}(F \leq f_{obs}) \leq \frac{\alpha}{2}$ ou si $P_{H_0}(F \geq f_{obs}) \leq \frac{\alpha}{2}$.

1.3.6 Test sur le coefficient de corrélation (avec $\sigma_1 = \sigma_2$ inconnus)

Le problème ici est de décider si une corrélation observée entre deux caractères statistiques, mesurée sur les mêmes individus, est ou non significative.

Les observations proviennent d'un échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ d'une loi normale bidimensionnelle, d'espérance (μ_x, μ_y) et de matrice de covariance :

$$\begin{pmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{pmatrix}.$$

C'est la loi d'un couple de variables, dont les espérances respectives sont μ_x et μ_y et les variances σ_x^2 et σ_y^2, le coefficient de corrélation étant ρ. L'estimateur naturel de ρ est le coefficient de corrélation empirique, à savoir la variable aléatoire r suivante :

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}},$$

où $\overline{X}$ et $\overline{Y}$ désignent les moyennes empiriques des X_i et des Y_i respectivement.

On suppose que l'échantillon aléatoire composé de deux variables est un vecteur Gaussien dont les variables suivent respectivement des lois $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$.

- Considérons les trois problèmes de test suivants :

$a)$ $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$ (test unilatéral à droite)

$b)$ $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{cases}$ (test unilatéral à gauche)

$c)$ $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$ (test bilatéral)

• La statistique du test est donnée par :

$$T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

Sous H_0, $T \sim t_{n-2}$.

• Région critique : on rejette H_0 au niveau de signification α si

a) $t_{obs} \geq k_\alpha$ où k_α est défini par

$$P_{H_0}(T \geq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \geq k_\alpha\}.$$

b) $t_{obs} \leq k_\alpha$ où k_α est défini par

$$P_{H_0}(T \leq k_\alpha) = \alpha.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq k_\alpha\}.$$

c) $t_{obs} \leq k_{1,\alpha}$ ou si $t_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(T \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(T \geq k_{2,\alpha}) = \frac{\alpha}{2}.$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq k_{1,\alpha} \text{ ou } t \geq k_{2,\alpha}\}.$$

• Calcul de la p-valeur : on rejette H_0, au niveau de signification α, si

a) $P_{H_0}(T \geq t_{obs}) \leq \alpha$

b) $P_{H_0}(T \leq t_{obs}) \leq \alpha$

c) $P_{H_0}(T \leq t_{obs}) \leq \frac{\alpha}{2}$ ou si $P_{H_0}(T \geq t_{obs}) \leq \frac{\alpha}{2}$.

Chapitre 2

tests de permutations

2.1 Introduction

La méthode de tests de permutations est une approche très générale des tests d'hypothèse statistique. Au lieu de comparer la vraie valeur de la statistique du test avec une distribution standard, la distribution de référence est générée par les données elles-mêmes, en permutant des données d'échantillons.

Pour la plupart des tests d'hypothèse classiques, nous recherchons d'abord, à partir des hypothèses, la distribution d'échantillonnage de la statistique du test sous l'hypothèse nulle H_0. Pour les tests de permutations nous renverserons la procédure, puisque la distribution d'échantillonnage est donnée par les permutations.

2.1.1 Origine des tests de permutations

L'agronome R. A. Fisher proposa dès 1935 la méthode des tests par permutations. Néanmoins, cette méthode nécessite des milliers de calculs et comme les statisticiens de l'époque ne disposaient pas d'ordinateurs, ils furent obligés de trouver une autre méthode bien plus rapide à réaliser. C'est pourquoi ils développèrent les différentes distributions de probabilités qui servent plus couramment à réaliser les tests statistiques. Ceux-ci ont réussi à démontrer, avec beaucoup d'ingéniosité, que sous certaines conditions, certaines statistiques du test obéissent à certaines lois particulières lorsque H_0 est vraie.

Par exemple :

• La variable-test t, utilisée dans le test de comparaison des moyennes de deux échantillons, obéit à une loi de Student (munie d'un certain nombre de degrés de liberté), si H_0 est vraie.

• La variable-test F, utilisée dans le test de comparaison des moyennes de plusieurs groupes (analyse de variance), obéit à une loi de F (munie de certains degrés de liberté), si H_0 est vraie.

• La variable-test t, utilisée dans le test des coefficients de corrélation, obéit à une loi de Student (munie d'un certain nombre de degrés de liberté), si H_0 est vraie.

Grâce à ce subterfuge, il devient possible d'employer une loi théorique au lieu de générer soi-même une loi empirique de la variable auxiliaire sous H_0, ce qui épargne beaucoup de travail.

Mais aujourd'hui, grâce à la vitesse des ordinateurs modernes, nous pouvons pratiquer des tests statistiques en utilisant la méthode des permutations. L'avantage principal est que nous ne devons pas nous soucier des hypothèses restrictives des distributions des procédures de test classique. Par contre, le désavantage est le temps de calcul nécessaire pour faire un grand nombre de permutations, chacune d'elles étant suivie d'un nouveau calcul de la statistique du test. Néanmoins ce désavantage disparaît au fur et à mesure que la puissance des ordinateurs augmente.

2.1.2 Permutations

Rappelons que mathématiquement une permutation de n éléments est un "réordonnement" de ceux-ci.

Définition : Une permutation des nombres entiers compris entre 1 et k est une bijection de l'ensemble {1, ... , k} dans lui-même.

On peut donc dire qu'une permutation des entiers compris entre 1 et k revient à considérer ces entiers dans un ordre nouveau.

Par exemple,

$$(1, 2, 3, 4, 5)$$
$$(1, 3, 2, 4, 5)$$
$$(4, 5, 2, 1, 3)$$
$$(3, 2, 1, 4, 5)$$
$$\ldots$$

sont toutes les permutations des numéros 1 à 5 (notons que ceci inclut la première ligne standard sans changement).

Il y a $n!$ permutations différentes possibles d'entiers de 1 à n objets. Dans ce cas-ci, $5! = 120$.

Remarquons également que si nous effectuons des permutations entre deux ensembles contenant respectivement n_1 et n_2 éléments, il n'y aura que $\frac{(n_1+n_2)!}{n_1!n_2!}$ permutations distinctes.

2.1.3 Distribution de la statistique du test

Au lieu de comparer la vraie valeur de la statistique avec une distribution standard, la distribution de référence est générée par les données elles-mêmes.

L'argument invoqué pour construire une distribution sous l'hypothèse nulle pour la statistique est que si l'hypothèse nulle H_0 est vraie, tous les éléments du premier échantillon ont quasiment la même chance d'apparaître dans le second et inversement. L'agencement des éléments entre les deux échantillons est dû seulement à la chance ; c'est-à-dire que n'importe quelle valeur du premier échantillon peut être échangée avec n'importe quelle valeur de second échantillon.
Les échantillons des données observées sont juste une des permutations possibles des éléments des deux échantillons, c'est pourquoi la valeur de la statistique du test pour des données non permutées (s_{obs}) devraient être typiques, c'est-à-dire située dans une partie centrale de la distribution des permutations.

Une réalisation de H_0 est obtenue en permutant les valeurs entre les deux groupes. Pour les valeurs ainsi permutées, on calcule une valeur qu'aurait pu prendre la statistique du test, en supposant que H_0 soit vraie. En répétant cette opération un grand nombre de fois, les différentes permutations produisent un ensemble de valeurs de la statistique obtenue sous H_0 (S_{per}), qui

permettent une estimation de la distribution d'échantillonnage de la statistique sous H_0.

Il faut ajouter à celles-ci la valeur de référence de la statistique s_{obs}, calculée à partir des éléments des échantillons non permutés. Puisque H_0 est testée, cette valeur est considérée comme une valeur qui pourrait être obtenue sous H_0, et par conséquent, devrait appartenir à la distribution de référence. Ensemble, les valeurs permutées et non permutées forment une estimation de la distribution d'échantillonnage de la statistique du test S sous H_0.

Nous avons vu dans la section 3 que sous certaines conditions, les statistiques du test obéissent à certaines lois, lorsque H_0 est vraie. Ces conditions dépendent des tests. Dans de nombreux tests, par exemple, la loi de distribution ne s'applique au cas où H_0 est vraie que si les données sont extraites d'une population dont la distribution est Normale. Lorsque des données ne sont pas conformes aux hypothèses de distribution de la méthode statistique que l'on veut utiliser (par exemple : normalité), le test de permutations nous fourni une approche efficace pour faire un test d'hypothèse sur ces données. De plus, un test de permutations est applicable à de très petits échantillons, comme le sont par ailleurs certains tests non paramétriques.

Il ne résout cependant pas les problèmes d'indépendance des observations. La méthode ne résout pas non plus des problèmes de distribution qui sont liés aux hypothèses soumises par un test. Comme par exemple dans le cas de l'analyse de la variance ou le test t de la différence entre groupes, la distribution de référence ne peut strictement être employée que si les variances des différents groupes sont égales (condition d'égalité des variances, ou homoscédasticité). Cette condition concernera également les tests par permutation.

Ce qu'il faut retenir c'est que certaines des conditions d'application particulières à chaque test résultent de notre référence aux lois de distribution ; ces conditions ne sont donc pas inhérentes aux tests statistiques eux-mêmes. On peut passer outre aux conditions qui concernent la distribution des données, en général la normalité, en utilisant la méthode des permutations exposée plus haut.

C'est pour cela que le test de permutations reste la méthode choisie pour tester des statistiques originales ou autres dont la distribution n'est pas connue. De plus, les résultats de la méthode des permutations restent valables même avec des observations qui ne sont pas des échantillons aléatoires d'une certaine population.

2.1.4 Concept d'échangeabilité

Un test de permutations construit une distribution sous l'hypothèse nulle, par le fait que tous les éléments du premier échantillon ont quasiment la même chance d'apparaître dans le second et inversement. C'est-à-dire que la distribution des données sous l'hypothèse nulle satisfait la condition "d'échangeabilité".

Définition : Les observations $\{x_1, x_2, ..., x_n\}$ sont échangeables si la probabilité de n'importe quel résultat particulier commun à l'échantillon, est la même quelque soit l'ordre dans lequel les observations sont considérées.

Autrement dit, si les étiquettes identifiant les quantités individuelles aléatoires sont non informatives, dans le sens où l'information fournie par les x_i est indépendante de l'ordre dans lequel elles sont collectées.
Des observations indépendantes et identiquement distribuées sont échangeables.

"L'échangeabilité" des observations dans les échantillons combinés est une condition suffisante pour qu'un test de permutations soit exact et non biaisé (Lehmann, 1986, p231).

2.1.5 Décision statistique

Notons tout d'abord que s_{obs} représente la valeur de la statistique du test à partir des échantillons non permutés et S_{per} les valeurs de la statistique du test calculées après permutations des éléments des deux échantillons.

Comme dans n'importe quel autre test statistique, la décision se prendra en comparant la valeur de référence de la statistique du test (s_{obs}) à la distribution de référence obtenue sous H_0 par les différentes valeurs S_{per}. Si la valeur de référence de s_{obs} est caractéristique des valeurs obtenues sous l'hypothèse nulle, H_0 ne peut pas être rejetée. Par contre si la valeur de référence est inhabituelle, étant trop "extrême" pour être considérée comme un résultat probable sous H_0, H_0 est rejetée et l'hypothèse alternative est considérée comme étant une explication plus probable des données.

où P_{calc} est définie en fonction du type de test.

- Pour un test bilatéral :

$$P_{calc} = \frac{[S_{per} < -|s_{obs}|] + [S_{per} = -|s_{obs}|] + [S_{per} = |s_{obs}|] + [S_{per} > |s_{obs}|]}{\text{Nbre de permutations } + 1}$$

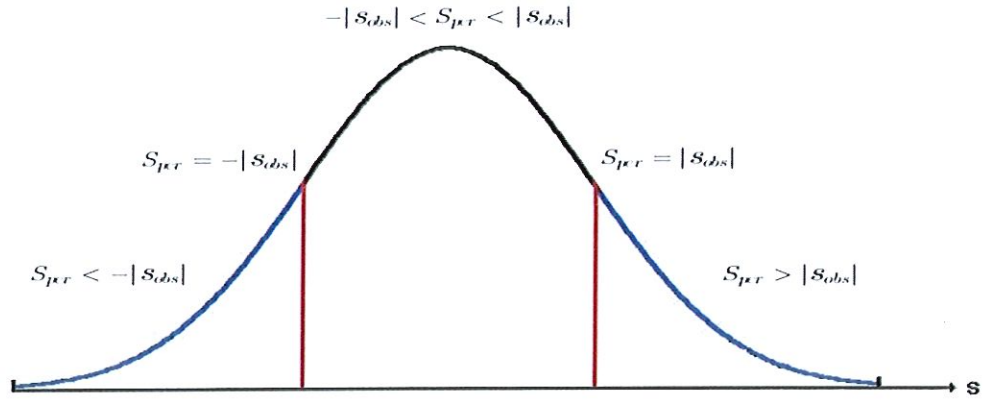
- Pour un test unilatéral à droite :

$$P_{calc} = \frac{[S_{per} = |s_{obs}|] + [S_{per} > |s_{obs}|]}{\text{Nbre de permutations } + 1}$$

- Pour un test unilatéral à gauche :

$$P_{calc} = \frac{[S_{per} < -|s_{obs}|] + [S_{per} = -|s_{obs}|]}{\text{Nbre de permutations } + 1}$$

Remarquons que le "+1" contenu dans le dénominateur vient du fait que l'on prend en compte la permutation nulle qui a permis de calculer la valeur observée à partir des deux échantillons non permutés.



Nous voyons à travers ce schéma ce que représente les différents intervalles $S_{per} < -|s_{obs}|$, $S_{per} = -|s_{obs}|$, $-|s_{obs}| < S_{per} < |s_{obs}|$, $S_{per} = |s_{obs}|$ et $S_{per} > |s_{obs}|$.

$S_{per} > |s_{obs}|$.

Pour un test bilatéral, par exemple, si la valeur observée s_{obs} est si élevée qu'elle est plus grande que la plupart des valeurs obtenues en simulant l'hypothèse H_0, c'est-à-dire que le nombre de S_{per} telles que $S_{per} \leq |s_{obs}|$ est très grand, ou encore si elle est si "fortement négative" que peu des valeurs obtenues en simulant H_0 sont aussi "fortement négatives", c'est-à-dire que le nombre de S_{per} telles que $S_{per} \leq -|s_{obs}|$ est très grand, alors on ne peut pas croire que les résultats expérimentaux sont compatibles avec l'hypothèse nulle et on rejette H_0.

Si, au contraire, la valeur observée s_{obs} se trouve vers le centre de la distribution des valeurs obtenues sous H_0, cela montre qu'une telle valeur aurait très bien pu être obtenue au hasard de l'échantillonnage d'une population ayant subi un traitement unique et on accepte donc que les données ne sont pas incompatibles avec l'hypothèse principale.

Si on se rappelle que le niveau de signification α d'un test d'une statistique est la proportion de valeurs qui sont aussi extrêmes ou plus extrêmes que la statistique du test dans la distribution de référence, on comprend dès lors pourquoi on rejette H_0 au niveau de signification α lorsque $P_{calc} \leq \alpha$.

2.2 Exemple

(Cet exemple a un but pédagogique afin de bien visualiser comment se déroule un test de permutations. Il n'a aucune prétention de vouloir tirer de vraies conclusions statistiques à partir de si petits échantillons.)

Rappelons brièvement la marche à suivre pour effectuer un test de permutations.

1. Analyser le problème.
2. Choisir une statistique de test.
3. Permuter des éléments des échantillons et recalculer la statistique du test afin d'obtenir une estimation de la distribution d'échantillonnage de cette statistique sous H_0.
4. Rejeter ou non l'hypothèse nulle.

Soient deux ensembles de 3 valeurs observées à partir de 2 groupes :

$$\begin{cases} \text{Groupe 1} = \{60, 63, 65\} \\ \\ \text{Groupe 2} = \{17, 27, 41\} \end{cases}$$

Supposons que nous voulions utiliser un test de permutations afin d'analyser la différence de moyenne entre les 2 groupes.
On prendra $\alpha = 0.05$ comme niveau de signification.

H_0 : il n'y a pas de différence entre la moyenne du groupe 1 et la moyenne du groupe 2.

H_1 : il a une différence entre la moyenne du groupe 1 et la moyenne du groupe 2 telle que la moyenne du groupe 1 est plus grande que celle du groupe2 (test unilatéral à droite).

On choisira comme statistique du test simplement la différence entre les moyennes des 2 échantillons : $S = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$.
Pour les échantillons non permutés i.e. $(60, 63, 65)$ et $(17, 27, 41)$, la valeur de la statistique du test :

$$s_{obs} = \frac{103}{3} .$$

Dans cet exemple, si les deux groupes ont la même moyenne alors les mouvements aléatoires des observations parmi les groupes ne produiront que de petites variations sur les valeurs de S_{per}. Certaines des S_{per} seront un peu plus grandes que 103/3, et d'autres plus petites.

Remarquons que si nous effectuons des permutations entre ces deux ensembles,
il n'y aura que $\frac{(n_1 + n_2)!}{n_1! n_2!}$ permutations distinctes.

Pour 6 observations, il y a 720 permutations possibles dont 20 combinaisons sont distinctes pour lesquelles on calculera S_{per}.

Voici donc un tableau contenant les 20 permutations distinctes ainsi que la valeur de la S_{per} associée :

Per	Groupe 1	Group 2	S_{per}
1	60 63 65	17 27 41	$\frac{103}{3}$
2	17 63 65	60 27 41	$\frac{17}{3}$
3	27 63 65	17 60 41	$\frac{37}{3}$
4	41 63 65	17 27 60	$\frac{65}{3}$
5	60 17 65	63 27 41	$\frac{11}{3}$
6	60 27 65	17 63 41	$\frac{31}{3}$
7	60 41 65	17 27 63	$\frac{59}{3}$
8	60 63 17	65 27 41	$\frac{7}{3}$
9	60 63 27	17 65 41	$\frac{27}{3}$
10	60 63 41	17 27 65	$\frac{55}{3}$

Per	Groupe 1	Group 2	S_{per}
11	17 27 65	60 63 41	$\frac{-55}{3}$
12	17 41 65	60 27 63	$\frac{-27}{3}$
13	27 41 65	17 63 60	$\frac{-7}{3}$
14	17 63 41	60 27 65	$\frac{-31}{3}$
15	17 63 27	60 65 41	$\frac{-59}{3}$
16	27 63 41	17 60 65	$\frac{-11}{3}$
17	60 17 27	63 65 41	$\frac{-65}{3}$
18	60 17 41	63 27 65	$\frac{-37}{3}$
19	60 27 41	17 63 65	$\frac{-17}{3}$
20	17 27 41	60 63 65	$\frac{-103}{3}$

Sur ces 20 "réordonnements" différents seulement deux ont une S_{per} plus grande ou égale que 103. De plus la probabilité que S_{per} soit plus grand ou égal à 103/3 est de $P_{calc} = 2/20 = 0.1 > 0.05 = \alpha$. On rejettera donc H_0 au niveau de signification $\alpha = 0.05$.

Dans ce cas les permutations ont donné lieu à un test exact car on a pu énumérer toutes les combinaisons possibles. Avec un exemple contenant bien plus de données, il aurait été impossible de lister toutes les permutations, même avec les ordinateurs modernes. On peut cependant obtenir des résultats approximatifs tout à fait valables en réalisant seulement quelques milliers de permutations, choisies de façon aléatoire parmi toutes les permutations possibles (habituellement de 1000 à 10000).

2.3 Remarques sur les tests de permutations

Dans les tests de permutations, la distribution de référence de la statistique du test est obtenue par des permutations aléatoires des données étudiées, sans référence à aucune population statistique. Le test est valable aussi longtemps que l'on génère la distribution de référence par la procédure décrite avec une hypothèse nulle qui donne du sens au problème, sans se soucier si les données sont, ou non, représentatives d'une plus grande population statistique. C'est la raison pour laquelle les données ne doivent pas être nécessairement un échantillon aléatoire simple d'une population statistique plus grande. La seule information que fournit le test de permutations est de dire si un modèle observé dans les données est probable ou non, d'être apparu par chance. Pour cette raison, certains pensent que les tests de permutations ne sont pas aussi bons ou intéressants que les tests classiques parce qu'ils ne nous autorisent pas à déduire des conclusions que l'on peut appliquer à une population statistique.

Un vue plus pragmatique des choses amène à la conclusion que la méthode des permutations peut être généralisée à une population de référence si le jeu de données est un échantillon aléatoire de cette population.

La généralisation des résultats, tant avec les tests classiques qu'avec les tests de permutations, dépend du caractère aléatoires des données.

2.3.1 test de permutations complet - partiel

Pour de petits ensembles de données, on peut calculer toutes les permutations possibles et obtenir ainsi la distribution de la statistique du test avec l'ensemble de toutes les permutations ; on a dès lors un test de permutations exact ou complet.

Pour les grands jeux de données, on ne peut calculer qu'une partie de toutes les permutations possibles. Lorsque l'on a un test de permutations partiel, il est important d'être sûr que celui-ci utilise un algorithme qui génère **uniformément** et **aléatoirement** les permutations c'est-à-dire capable de produire toutes les permutations possibles avec une probabilité égale.

2.3.2 Nombre de permutations

Le nombre de permutations que l'on considère dans un test partiel sera logiquement un compromis entre la précision requise d'une part et le temps mis par les ordinateurs pour effectuer les calculs d'autre part. Il est conseillé de faire le plus possible de permutations, puisque les estimations de probabilité sont sujettes à des erreurs dues à des permutations possibles de la population d'échantillonnage (sauf dans de rares cas de tests de permutations complets), mais il faut évidement prendre en compte le temps de calcul des ordinateurs lorsque l'on étudie de grands jeux de données.

La littérature traitant ce sujet (P.Legendre et L.Legendre) nous conseille pour une analyse exploratoire 1000 permutations. Si la probabilité calculée par ordinateur est proche du niveau de signification présélectionné, il faudra effectuer plus de permutations. Pour une analyse plus sérieuse, elle nous conseille d'utiliser au moins 10.000 permutations.

Néanmoins, comme la référence citée n'est pas très récente et que la puissance de calcul des ordinateurs a très fortement augmenté ces dernières années, nous allons comparer grâce à un exemple un test d'hypothèse avec respectivement avec 1.000, 10.000, 100.000 et enfin 1.000.000 de permutations afin de déterminer le nombre de permutations qui convient le mieux au problème. Notons que le test de permutations qui suit a été obtenu par un programme implémenté en *Matlab* sur un ordinateur muni d'un Intel Pentium 4 CPU 3.20Ghz avec 1 Go de ram.

Soient les deux échantillons suivants que l'on suppose indépendants :

Echantillon 1									
120	150	180	200	130	150	170	160	190	100
125	145	175	200	120	130	135	165	150	180

Echantillon 2									
115	118	135	185	195	170	155	180	191	200
100	98	105	135	145	155	118	120	112	130
118	125	135	155	165	156	187	198	127	130

Nous voulons savoir si les deux populations dont sont issus les échantillons ont des moyennes (respectivement μ_1 et μ_2) égales ou non. Nous effectuerons un test d'hypothèse avec un niveau de signification $\alpha = 0.05$.

Testons :

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 \\ H_1 &: \ \mu_1 \neq \mu_2 \end{aligned}$$

Le programme *Testdifmoyenneinconnue.m* a été implémenté pour réaliser ce test d'hypothèse.

On lui donne en entrée les éléments des deux échantillons ainsi que le nombre d'éléments des échantillons respectifs. Il utilise, comme statistique du test, la statistique t de student :

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}}}$$

où

$$S_j^2 = \frac{\sum_{i=1}^{n_j} \left(X_i - \overline{X}\right)^2}{n_j - 1}$$

avec $j \in \{1, 2\}$.

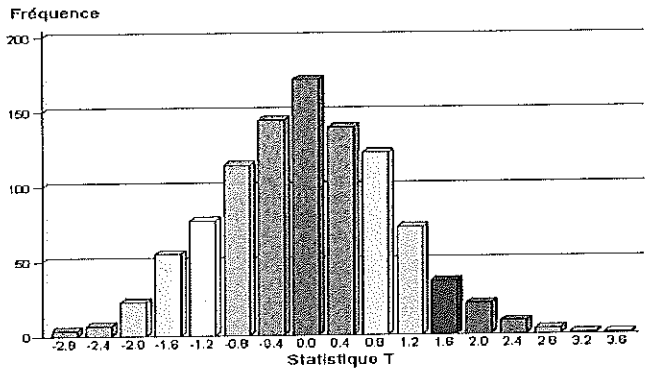
Il donne en sortie la valeur de la statistique à partir des échantillons non permutés s_{obs} ainsi que les variables *premier, deuxième, troisième, quatrième, cinquième* qui représentent respectivement le nombre de valeurs de la statistique S_{per} après permutation des échantillons telles que $S_{per} < -|s_{obs}|$ ou $S_{per} = -|s_{obs}|$ ou $-|s_{obs}| < S_{per} < |s_{obs}|$ ou $S_{per} = |s_{obs}|$ ou encore $S_{per} > |s_{obs}|$.

Le programme donne également le temps (en secondes) mis par le programme pour donner les résultats (*temps*). Il enregistre également dans un fichier toutes les données s_{per}.

Voici les résultats donnés par le programme *Testdifmoyenneinconnue.m* ainsi que l'histogramme associé de la distribution sous H_0 de la statistique T générée par les permutations. Cet histogramme a été réalisé grâce au logiciel *SAS*.

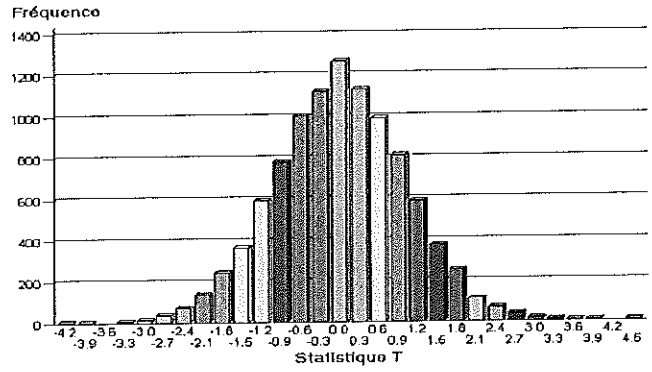
• Pour 999 permutations aléatoires :

s_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.9510	5.34	177	0	661	2	160



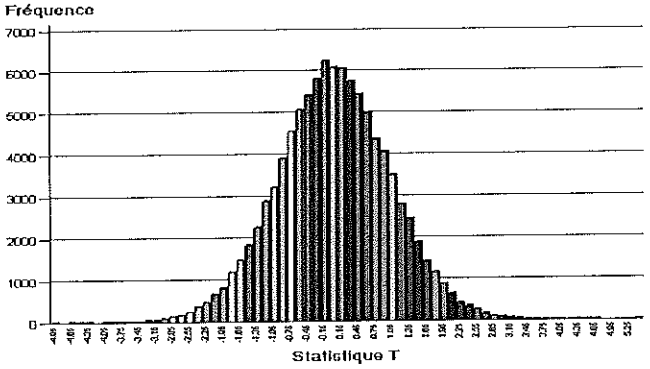
• Pour 9.999 permutations aléatoires :

s_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.9510	8.57	1671	0	6646	2	1681



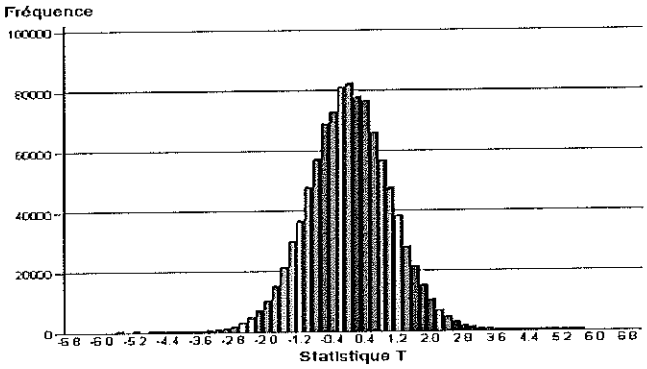
• Pour 99.999 permutations aléatoires :

s_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.9510	23.86	16637	0	66533	15	16815



• Pour 999.999 permutations aléatoires :

s_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.9510	172.80 $= 2'52''80$	167463	0	665875	228	166434



Nous constatons tout d'abord que, pour les 4 nombres de permutations ci-dessus, le test statistique mènera à la même conclusion ; on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$ puisque $P_{calc} > 0.05$ dans les 4 cas :

$$P_{calc1\,000} = \frac{[177] + [2] + [160]}{1000} = 0.339 > 0.05$$

$$P_{calc10\,000} = \frac{[1671] + [2] + [1681]}{10000} = 0.3354 > 0.05$$

$$P_{calc100\,000} = \frac{[16637] + [15] + [16815]}{10000} = 0.33467 > 0.05$$

$$P_{calc1\,000\,000} = \frac{[16637] + [15] + [16815]}{10000} = 0.334125 > 0.05$$

- Analyse de la précision

Lorsque l'on augmente le nombre de permutations d'un facteur dix, la précision des variables *premier, deuxième, troisième, quatrième, cinquième* augmente elle aussi d'un facteur dix. Comme le niveau de signification est donné généralement en pourcents, le résultat du P_{calc} d'une précision d'un millième suffira (pour l'arrondi). Or on constate, pour 99.999 et 999.999 permutations, que les 3 premiers chiffres après la virgule ne varient plus. C'est pour cette raison que nous délaisserons les tests de moins de 99.999 permutations qui ne sont pas assez précis.

Pour ce qui est des histogrammes qui représentent la distribution d'échantillonnage de la statistique sous H_0, les 999 permutations sont insuffisantes. Pour 9.999 permutations, l'histogramme commence à bien approcher une loi de distribution théorique. Pour 99.999 et 999.999, on voit clairement une distribution d'échantillonnage de la statistique sous H_0. on peut même dire que faire un million de permutation n'apporte pas grand chose en plus aux 99.999 permutations aléatoires.

• Point du vue temps de calcul

Le temps de calcul entre 999 et 9.999 permutations est quasi identique alors que celui pour 999.999 explose par rapport à celui de 99.999.

Comme on l'a déjà dit auparavant, le nombre de permutations que l'on choisira sera un compromis entre les temps de calcul et la précision. Tous les temps recueillis sont raisonnables, mais on remarque que l'on n'obtient pas beaucoup plus d'informations si l'on augmente le nombre de permutations de 99.999 à 999.999. Mais le fait de manipuler un million de résultats pour représenter l'histogramme de la distribution statistique approchée sous H_0 est très fastidieuse. C'est pour cette raison que dans ce mémoire, on choisira de faire des tests d'hypothèse avec un nombre de permutations aléatoires égale à 99.999.

Chapitre 3

Comparaison entre les tests classiques et les tests de permutations

Ce chapitre a pour but de confronter les tests d'hypothèse dit classiques aux tests par permutations et de montrer que l'on obtient des décisions statistiques similaires par ces deux approches.

Pour les raisons évoquées dans le chapitre précédent, nous effectuerons des tests d'hypothèse avec 100.000 permutations. Tous les résultats obtenus le seront à partir d'un Intel Pentium 4 CPU 3.20 Ghz avec 1 Go de ram. Les différents programmes en *Matlab* utilisés sont joints dans les annexes du mémoire. Notons également que la variable *temps* est en secondes.

3.1 Tests sur la différence de deux moyennes avec σ_1 et σ_2 connus

3.1.1 Enoncé

Deux pisciculteurs élèvent des truites arc en ciel dans des installations quasi identiques. Le premier donne de la nourriture standard pour poisson tandis que le second donne des granulés conçus expressément pour truites arc en ciel garantissant selon le fournisseur les ressources nécessaires et permettant une croissance rapide et meilleure qu'à la normale. Evidement cette nourriture spécifique coûte plus cher que la nourriture standard.

Tous deux vendent leurs truites lorsqu'elles sont âgées d'un an et demi. Le deuxième pisciculteur se demande donc si ce surcoût lui donne de plus grandes truites.

FIG. 3.1 – Truite arc en ciel

Pour répondre à cette question, nous allons mesurer la longueur d'une partie des truites retirées du bassin 1 et du bassin 2 avant d'être vendues. Nous obtenons de cette façon deux échantillons aléatoires que l'on suppose indépendants :

Truites du bassin 1									
29	31	32	32	34	35	37	37	38	38
39	39	40	40	40	41	41	41	42	42
43	43	44	45	45	46	48	41	51	53

Truites du bassin 2									
27	30	30	31	31	34	35	36	36	37
38	40	40	41	41	42	42	43	43	43
44	45	47	47	47	49	52	54	56	58

Les populations, desquels les échantillons sont issus, suivent respectivement des lois $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$ (vérification via le programme *chi2test.m*). De plus, grâce à de nombreux contrôles de qualité effectués lors de la vente de truites, nous savons que $\sigma_1 = 30$ cm et $\sigma_2 = 60$ cm.

3.1.2 Hypothèse nulle

Nous allons donc tester si la longueur moyenne des truites du bassin 2 est plus grande ou égale à celle du bassin 1. Il s'agira donc d'un test unilatéral à gauche.
On émettra donc les hypothèses suivantes :

H_0 : Il n'y a pas de différence entre les longueurs moyennes des truites des deux bassins.

$\mu_1 = \mu_2$ ou encore $\mu_1 - \mu_2 = 0$.

H_1 : La longueur moyenne des truites du bassin 2 est plus grande que celle des truites du bassin 1.

$\mu_1 < \mu_2$ ou encore $\mu_1 - \mu_2 < 0$.

On fixe le niveau de signification α à 0.05 .

3.1.3 Statistique du test

On choisira comme statistique du test, la statistique Z :

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} .$$

3.1.4 Test classique

Région critique :

On rejette H_0, au niveau de signification α, si $z_{obs} \leq k_\alpha$ où k_α est défini par

$$P_{H_0}(Z \leq k_\alpha) = \alpha \text{ . On trouve } k_{0.05} = -1.645 \text{ .}$$

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, z \leq -1.645\} \text{ .}$$

Or $z_{obs} = -0.6158 > -1.645$.

On ne rejette donc pas H_0 au niveau de signification $\alpha = 0.05$.

p-valeur

On rejette H_0 au niveau de signification α, si :

$$P_{H_0}(Z \leq z_{obs}) \leq \alpha \text{ .}$$

En regardant dans la tables de la loi Normale centrée réduite, on voit que

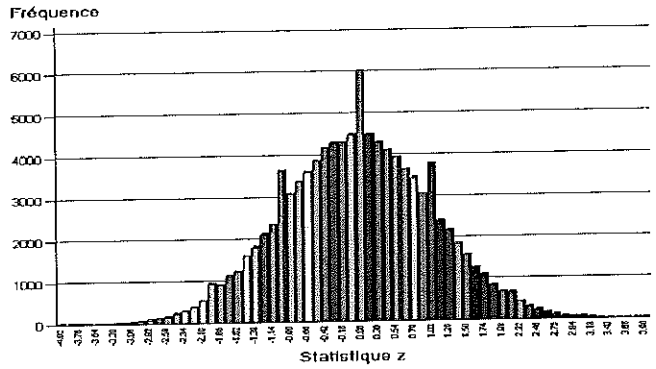
$$P_{H_0}(Z \leq -0.6158) = 0.269013 > 0.05$$

et on ne rejettera donc pas H_0 au seuil de signification $\alpha = 0.05$.

3.1.5 Test par permutations

Le programme *Testdifmoyenneconnue.m* donne, pour 99.999 permutations aléatoires, les résultats suivants :

z_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
−0.6158	14.21	26663	1227	44044	1256	26810



$$P_{calc} = \frac{[26663] + [1227]}{100000} = 0.2789 > 0.05$$

La régle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

3.1.6 Conclusion

Les deux méthodes donnent des résultats équivalents et le pisciculteur n'a pas intérêt à nourrir ses poissons avec de la nourriture spéciale puisqu'aucune différence notable entre la longueur moyenne des truites des deux bassins n'a été décelée (au niveau de signification 0.05).

3.2 Tests sur la différence de deux moyennes avec $\sigma_1 = \sigma_2$ inconnus

3.2.1 Enoncé

Un chercheur a effectué des relevés sur la taille des Narcisses lors de ses recherches aux îles Canaries et en Espagne sur le bassin méditerranéen. Il voudrait savoir si les Narcisses qu'il a vus aux îles Canaries et sur le continent sont originaires de la même espèce.

FIG. 3.2 – Narcisses

Comme la hauteur de ces plantes peut déterminer l'espèce dont elles sont issues, le chercheur va comparer la hauteur moyenne de ces plantes à partir des mesures prises lors de ses voyages. Il a prélevé sur le terrain deux échantillons aléatoires de 25 mesures de la taille des fleurs respectivement aux Canaries et en Espagne. Les échantillons sont supposés indépendants. Les résultats sont les suivants :

Narcisses des Canaries									
149	150	152	153	153	155	156	157	158	158
158	158	159	159	160	162	162	162	163	163
164	164	165	171	174					

Narcisses du continent									
140	150	151	152	153	154	155	156	157	158
158	158	158	159	160	160	162	162	162	163
164	164	165	166	171					

Les populations desquelles ces échantillons aléatoires ont été tirés, suivent respectivement des lois $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$. La vérification de la normalité des données a été effectuée à l'aide du programme *chi2test.m*.
σ_1 et σ_2 sont inconnus et on suppose que $\sigma_1 = \sigma_2$. On vérifiera cette hypothèse dans l'exemple suivant.

3.2.2 Hypothèse nulle

Le chercheur veut tester si la hauteur moyenne des Narcisses d'Espagne et des Canaries sont identiques, ou non. Il s'agira donc d'un test bilatéral. On émettra donc les hypothèses suivantes :

H_0 : Il n'y a pas de différence entre la hauteur moyenne des Narcisses provenant d'Espagne ou des Canaries.

$\mu_1 = \mu_2$ ou encore $\mu_1 - \mu_2 = 0$.

H_1 : Il existe une différence entre la hauteur moyenne des Narcisses selon qu'ils sont issues de l'Espagne ou des Canaries.

$\mu_1 \neq \mu_2$ ou encore $\mu_1 - \mu_2 \neq 0$.

On fixe le niveau de signification α à 0.05 .

3.2.3 Statistique du test

On choisira comme statistique du test, la statistique T :

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

3.2.4 Test classique

Région critique :

On rejette H_0, au niveau de signification α, si $t_{obs} \leq k_{1,\alpha}$ ou si $t_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(T_{48} \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(T_{48} \geq k_{2,\alpha}) = \frac{\alpha}{2} .$$

On trouve après calcul $k_{1,0.05} = -2.0106$ et $k_{2,0.05} = 2.0106$.
La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq -2.0106 \text{ ou } t \geq 2.0106\} .$$

Or $t_{obs} = 0.6097$ et donc $-2.0106 < 0.6097 < 2.0106$.

On ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

p-valeur

On rejette H_0, au niveau de signification α, si :

$$P_{H_0}(T_{48} \leq t_{obs}) \leq \frac{\alpha}{2} \quad \text{ou si} \quad P_{H_0}(T_{48} \geq t_{obs}) \leq \frac{\alpha}{2} .$$

En regardant dans la table de la loi de Student, on voit que

$$P_{H_0}(T_{48} \leq 0.6097) = 0.5449 > 0.025$$

et

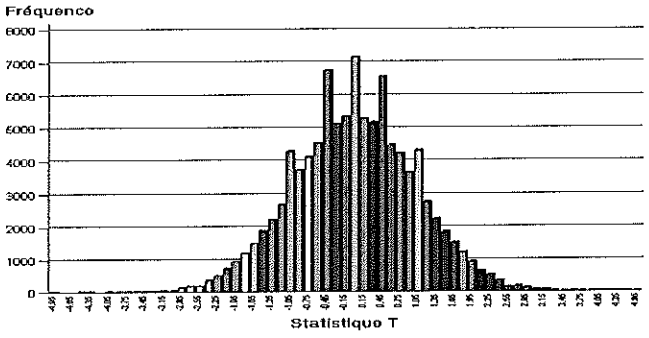
$$P_{H_0}(T_{48} \geq 0.6097) = 1 - 0.5449 = 0.4551 > 0.025 .$$

On ne rejette donc pas H_0 au seuil de signification $\alpha = 0.05$.

3.2.5 Test par permutations

Le programme *Testdifmoyenneinconnue.m* donne les résultats suivants pour un nombre de 99.999 permutations aléatoires :

t_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.6097	21.66	26375	1360	44515	1318	26432



$$P_{calc} = \frac{[26375] + [1360] + [1318] + [26432]}{100000} = 0.55485 > 0.05$$

La régle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

3.2.6 Conclusion

Les deux méthodes conduisent à la même conclusion (au niveau de signification $\alpha = 0.05$). Le chercheur peut penser que les fleurs de la péninsule ibérique et des Canaries appartiennent à la même espèce puisque aucune différence notable entre les hauteurs moyennes des Narcisses des deux groupes n'a été décelée.

3.3 Tests sur le rapport de deux variances

Dans l'exemple précédent, pour pouvoir utiliser le test t de la différence entre groupes, la distribution de référence ne peut strictement être employée que si les variances des deux groupes sont égales. Nous allons donc montrer que cette condition est bien remplie et que nous pouvions donc utiliser la statistique t. Ce test va être effectué, d'une part de manière classique, et d'autre part par les tests de permutations.

3.3.1 Enoncé

Nous reprenons donc exactement les mêmes données concernant les Narcisses de l'exemple précédent.

3.3.2 Hypothèse nulle

Le chercheur veut tester si la variance de la hauteur des Narcisses des Canaries est égale à celle des Narcisses d'Espagne ou non. Il s'agira donc d'un test bilatéral.

On émettra donc les hypothèses suivantes :

H_0 : Il n'y a pas de différence entre la variance de la hauteur des Narcisses selon qu'elles sont issues d'Espagne ou des Canaries.

$$\sigma_1^2 = \sigma_2^2$$

H_1 : La variance de la hauteur des Narcisses des Canaries est différente de la variance de la hauteur des Narcisses d'Espagne.

$$\sigma_1^2 \neq \sigma_2^2$$

On fixe le niveau de signification α à 0.05 .

3.3.3 Statistique du test

On choisira comme statistique du test, la statistique F :

$$F = \frac{\frac{(n_1 - 1)}{\sigma_1^2} S_1^2}{\frac{(n_2 - 1)}{\sigma_2^2} S_2^2}$$

où $S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_i - \overline{X}\right)^2$, $j \in \{1, 2\}$.

3.3.4 Test classique

Région critique :

On rejette H_0, au niveau de signification α, si $f_{obs} \leq k_{1,\alpha}$ ou si $f_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(F_{25,25} \leq k_{1,\alpha}) = \frac{\alpha}{2} \quad \text{et} \quad P_{H_0}(F_{25,25} \geq k_{2,\alpha}) = \frac{\alpha}{2}.$$

On trouve $k_{1,0.05} = 0.4484$ et $k_{2,0.05} = 2.2303$ et la région critique est donnée par

$$W = \{(x_1, ..., x_n) \mid f \leq k_{1,\alpha} \text{ ou } f \geq k_{2,\alpha}\}.$$

Or $f_{obs} = 0.8799$ et $0.4484 < 0.8799 < 2.2303$.

On ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

p-valeur

On rejette H_0 au niveau de signification α, si :

$$P_{H_0}(F_{25,25} \leq f_{obs}) \leq \frac{\alpha}{2} \text{ ou si } P_{H_0}(F_{25,25} \geq f_{obs}) \leq \frac{\alpha}{2}.$$

En regardant dans la table de la loi de Fisher, on voit que

$$P_{H_0}(F_{25,25} \leq 0.8799) = 1 - 0.6242 > 0.025$$

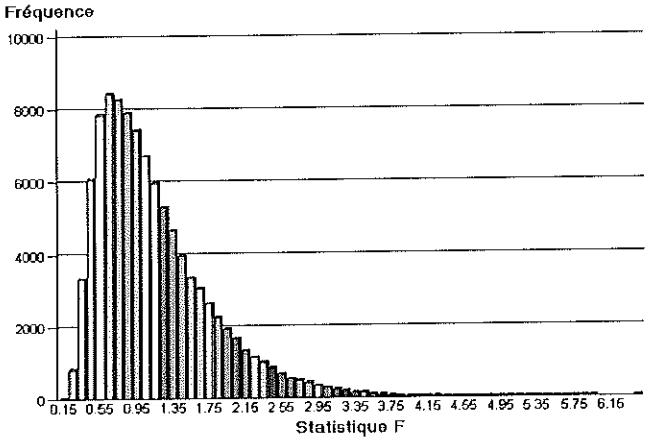
$$P_{H_0}(F_{25,25} \geq 0.8799) = 0.6242 > 0.025 .$$

On ne rejette pas H_0 au seuil de signification $\alpha = 0.05$.

3.3.5 Test par permutations

Le programme *Testrapvariance.m* donne pour un nombre de 99.999 permutations aléatoires les résultats suivants :

f_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
0.8799	22.08	0	0	40987	152	58861



$$P_{calc} = \frac{[152] + [58861]}{100000} = 0.59013 > 0.05 \, .$$

La régle de décision est la suivante : on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

3.3.6 Conclusion

Encore une fois les deux types de test donnent les mêmes décisions statistiques ; on pouvait donc utiliser le test de la section 2 pour mesurer la différence de deux moyennes avec $\sigma_1 = \sigma_2$ inconnus.

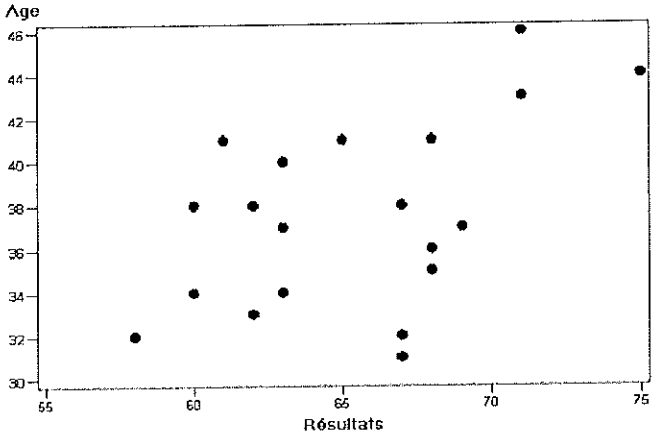
3.4 Tests de corrélation

3.4.1 Enoncé

Une entreprise souhaite engager des employés et pour cela elle veut utiliser un nouveau test coté sur 100. Ce test comporte de nombreuses questions de culture générale. La cellule qui s'occupe du recrutement se demande si ce nouveau type de test n'est pas faussé en fonction de l'âge de la personne questionnée.

Avant de valider ce nouveau type de test, l'entreprise fait passer celui-ci à ses employés. Nous collectons ainsi sur 20 individus des informations concernant les deux variables que l'on veut confronter : l'âge et le résultat du test.

Voici le diagramme de dispersion et les données elles-mêmes :



Individus	Résultats	Age
1	61	41
2	71	46
3	62	38
4	75	44
5	58	32
6	60	38
7	67	31
8	68	41
9	71	43
10	69	37

Individus	Résultats	Age
11	68	35
12	67	32
13	63	37
14	62	33
15	60	34
16	63	40
17	65	41
18	67	38
19	63	34
20	68	36

On considère ces données comme un échantillon aléatoire d'une loi Normale bivariée dont les deux variables suivent respectivement des lois $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$, où $\sigma_1 = \sigma_2$ sont inconnus.

3.4.2 Hypothèse nulle

Nous voulons savoir s'il existe une relation entre l'âge et les résultats du test, ou non. Il s'agira donc d'un test bilatéral.
On émettra donc les hypothèses suivantes :

H_0 : Il n'y a pas de relation entre l'âge et les résultats du test.

$$\rho = 0$$

H_1 : Il existe une relation entre l'âge et les résultats du test.

$$\rho \neq 0$$

On fixe le niveau de signification α à 0.05 .

3.4.3 Statistique du test

On choisira comme statistique du test, la statistique T :

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\,.$$

3.4.4 Test classique

Région critique :

On rejette H_0 au niveau de signification α, si $t_{obs} \leq k_{1,\alpha}$ ou si $t_{obs} \geq k_{2,\alpha}$ où $k_{1,\alpha}$ et $k_{2,\alpha}$ sont définis par

$$P_{H_0}(T_{38} \leq k_{1,\alpha}) = \frac{\alpha}{2} \text{ et } P_{H_0}(T_{38} \geq k_{2,\alpha}) = \frac{\alpha}{2}$$

Après calcul on trouve $k_1 = -2.0244$ et $k_2 = 2.0244$.

La région critique est donnée par

$$W = \{(x_1, ..., x_n) \,|\, t \leq -2.0244 \text{ ou } t \geq 2.0244\}\,.$$

Or $t_{obs} = 2.4178$ et $2.0244 < 2.4178$. On rejette H_0 au niveau de signification $\alpha = 0.05$.

p-valeur

On rejette H_0 au niveau de signification α, si :

$$P_{H_0}(T \leq t_{obs}) \leq \frac{\alpha}{2} \text{ ou si } P_{H_0}(T \geq t_{obs}) \leq \frac{\alpha}{2} .$$

En regardant dans la table de la loi de Student, on voit que

$$P_{H_0}(T_{38} \leq 2.4178) = 0.0205 < 0.025$$

et

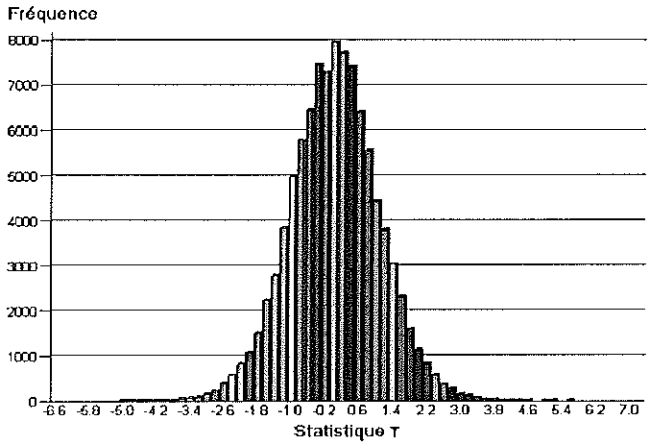
$$P_{H_0}(T_{38} \geq 2.4178) = 1 - 0.0205 > 0.025 .$$

On rejette H_0 au seuil de signification $\alpha = 0.05$.

3.4.5 Test par permutations

Le programme *Testcorrelation.m* donne pour 99.999 permutations aléatoires les résultats suivants :

t_{obs}	*temps*	premier	deuxième	troisième	quatrième	cinquième
2.4178	17.38	1290	0	97293	21	1396



$$P_{calc} = \frac{[1290] + [21] + [1396]}{100000} = 0.02707 < 0.05$$

La régle de décision est la suivante : on rejette H_0 au niveau de signification $\alpha = 0.05$.

3.4.6 Conclusion

Il existe donc bel et bien une relation entre l'âge et les résultats du nouveau test. Il faudra donc revoir celui-ci si on veut avoir une égalité des chances entre les jeunes demandeurs d'emploi et leurs aînés.

Chapitre 4

La méthode de classification des Hypervolumes

4.1 Le problème

L'objectif de la classification automatique est de trouver une éventuelle structure dans des jeux de données. Décomposer une population de données, d'individus ou d'objets, décrits par un ensemble de caractéristiques, en un certain nombre de groupes homogènes appelés **classes**.

Le problème de classification automatique peut se définir mathématiquement comme suit :

Soit $E = \{x_1, ..., x_n\}$, l'ensemble des individus à classifier. On suppose que $\#E < \infty$.

Soient $Y_1, ..., Y_p$, p variables qui caractérisent chaque objet.

L'objectif de la classification automatique est de :

- trouver k classes **naturelles** dans E ($k = ?$)
- valider les classes obtenues c'est-à-dire coller une "étiquette" sur chaque classe.

Ou plus précisément, on recherche une partition $P = \{C_1, ..., C_k\}$ de E en k classes (k fixé).

La plupart des méthodes de classification se basent sur un critère qui mesure la qualité de chaque partition P en k classes.

Soit $\mathcal{P}_k$ l'ensemble de toutes les partitions de E en k classes. Le critère de classification associé à chaque $P \in \mathcal{P}_k$ est :

$$W : \mathcal{P}_k \to \mathbb{R} \ : \ P \rightsquigarrow W(P,k) \ .$$

Le problème de classification revient alors à trouver la partition optimale $P^* = \{C_1^*, ..., C_k^*\}$ qui minimise la valeur du critère, c'est-à-dire telle que :

$$W(P^*,k) = \min_{P \in \mathcal{P}_k} W(P,k) \ .$$

4.2 La méthode des Hypervolumes

La méthode de classification automatique des Hypervolumes se différentie de la plupart des autres méthodes car le critère de celle-ci n'est pas basé sur des mesures de dissimilarités. En effet elle repose sur un modèle statistique dont l'approche est basée sur la théorie des processus ponctuels et sur l'estimation par la méthode du maximum de vraisemblance d'un domaine convexe compact.

4.2.1 Le modèle

Le modèle se base sur les hypothèses suivantes :

- Les points sont distribués dans k sous-domaines disjoints $D_1, D_2, ..., D_k$ de D.

- Les variables aléatoires qui comptent le nombre de points dans des régions disjointes sont indépendantes.

- Le nombre moyen de points dans chaque région est proportionnel à la mesure de Lebesgue de cette région.

Le seul processus qui satisfait à ces conditions est le processus de Poisson homogène que nous définissons ci-après.

4.2.2 Processus de Poisson homogène

N est un processus de Poisson homogène d'intensité q sur $D \subset \mathbb{R}^p$ $(0 < m(D) < \infty)$ si :

- $\forall A_1, ..., A_k \subset D$, $\forall i \neq j \in \{1, ..., k\}$ avec $A_i \cap A_j = \emptyset$, $N(A_i)$ est indépendant de $N(A_j)$ où $N(A_i)$ est la variable aléatoire qui dénombre A_i

- $\forall A \subset D$, $\forall k \geq 0$,

$$P\left(N(A) = k\right) = \frac{\left(q\, m(A)\right)^k}{k!} e^{-q\, m(A)}$$ où $m(\cdot)$ est la mesure de Lebesgue.

La variable aléatoire $N(A)$ suit donc une loi de Poisson dont la moyenne vaut $q\, m(A)$.

4.2.3 Propriété d'uniformité conditionnelle

Si $N(A) = n$ est fini , alors les n points sont distribués aléatoirement et uniformément dans A. Cette propriété d'uniformité conditionnelle nous permet d'écrire la fonction de densité associée au processus de Poisson homogène

$$f(x) = \frac{1}{m(D)} I_D(x)$$

où $m(D)$ est la mesure de Lebesgue de D, $I_D(x_i)$ la fonction indicatrice de D. Et donc, si $x = (x_1, ..., x_n)$, la fonction de vraisemblance peut s'écrire comme :

$$L_D(x) = \frac{1}{(m(D))^n} \prod_{i=1}^{n} I_D(x_i) \,.$$

4.2.4 Problème de base : l'estimation d'un ensemble convexe

On considère l'ensemble $E = \{x_1, ..., x_n\}$ des individus à classer; les points sont générés par un processus de Poisson homogène dans $D \in \mathbb{R}^p$ où

- $D = \bigcup_{i=1}^{k} D_i$, k fixé

- les ensembles D_i sont disjoints et convexes.

Le problème est d'estimer le paramètre D, c'est-à-dire les k domaines D_i disjoints et convexes en utilisant des méthodes d'inférence statistique. Pour cela, on recherche l'estimateur du maximum de vraisemblance de D, c'est-à-dire les estimateurs des domaines D_i qui maximisent la fonction de vraisemblance.

Or le domaine D pour lequel la vraisemblance est maximal est, parmi ceux qui contiennent tous les points, celui dont la mesure de Lebesgue est minimale.

On peut montrer (Ripley et Rasson (1977)) que si D est convexe, l'enveloppe convexe des points appartenant à D est une statistique exhaustive pour D. C'est aussi l'estimateur du maximum de vraisemblance de D.
Remarquons que cet estimateur est biaisé et que l'on peut résoudre ce problème en prenant comme estimateur de D la dilatation de l'enveloppe convexe à partir du centre de gravité des points de l'enveloppe convexe.

Une approximation du coefficient de dilatation est donnée par (Moore (1984))

$$c = \sqrt{\frac{n}{n - V_n}}$$

où V_n est le nombre de points sur l'enveloppe convexe.

Ainsi, le maximum de la fonction de vraisemblance sera atteint par la partition pour laquelle la somme des mesures de Lebesgue des enveloppes convexes des k sous-domaines est minimale.

4.2.5 Le critère

L'estimation du maximum de vraisemblance nous permet de déterminer le critère suivant ; on recherche une partition P de E en k classes (k fixé).

La méthode de classification des Hypervolumes suppose que les n points d'observation de dimension p, $\{x_1, ..., x_n\}$, sont générés par un processus de Poisson homogène N dans $D \subset \mathbb{R}^p$, où D est l'union de k domaines convexes disjoints $D_1, ...D_k$.
Le problème est d'estimer les domaines inconnus D_i dans lesquels les points ont été générés.

On note $C_i \subset \{x_1, ..., x_n\}$ le sous-ensemble de points appartenant à D_i $(1 \leq i \leq k)$, et $H(C_i)$ l'enveloppe convexe des points de C_i.

La fonction de vraisemblance est donnée par :

$$L(D; x_1, ..., x_n) = f(x_1, ..., x_n; D) = \prod_{i=1}^{n} f_{X_i}(x_i; D) = \prod_{i=1}^{n} \frac{1}{m(D)} I_D(x_i) .$$

Et donc

$$L(D; x_1, ..., x_n) = \frac{1}{(m(D))^n} \prod_{i=1}^{n} I_D(x_i) = \frac{1}{(m(D))^n} I_D\left(H(x_1, ..., x_n)\right)$$

où $H(x_1, ..., x_n)$ est l'enveloppe convexe de $(x_1, ..., x_n)$.

Maximiser la fonction de vraisemblance revient à minimiser le critère des Hypervolumes :

$$\max_{D_1,...,D_k} L_D(x) \Longleftrightarrow \min_{P \in \mathcal{P}_k} \sum_{i=1}^{k} m\left(H(C_i)\right) \Longleftrightarrow \min_{P \in \mathcal{P}_k} W_k .$$

Les estimateurs du maximum de vraisemblance des k domaines inconnus $D_1, ..., D_k$ sont les k enveloppes convexes $H(C_i)$ telles que

$$\sum_{i=1}^{k} m\left(H(C_i)\right) \text{ est minimale.}$$

Le critère des Hypervolumes sera donc défini par :

$$W_k : \mathcal{P}_k \longrightarrow \mathbb{R}^p : P \rightsquigarrow W(P, k) = \sum_{l=1}^{k} m\left(H(C_l)\right) = \sum_{l=1}^{k} \int_{H(C_l)} m(dx) .$$

La méthode de classification des Hypervolumes cherche à minimiser W_k sur l'ensemble des partitions en k classes. Dans le contexte de ce problème de classification, nous essayerons de trouver la partition P^* telle que

$$P^* = \arg \min_{P \in \mathcal{P}_k} \sum_{l=1}^{k} \int_{H(C_l)} m(dx).$$

Pratiquement, si le nombre de variable $p = 1$, la somme des mesures de Lebesgue des enveloppes convexes des classes sera la somme des longueurs

des intervalles qui constituent les classes. Si $p = 2$, la somme des mesures de Lebesgue sera la somme des aires des enveloppes convexes.

4.2.6 Processus de Poisson non homogène

N est un processus de Poisson non homogène d'intensité q sur $D \subset \mathbb{R}^p$ $(0 < m(D) < \infty)$ si :

- $\forall A \subset D$, $N(A)$ a une distribution de Poisson de paramètre $\int_A q(x)\, m\,(dx)$.
- Propriété conditionnelle : si $N(A) = n$, alors les n points sont distribués indépendamment dans A, avec une fonction de densité proportionnelle à $q(x)$.

4.2.7 La méthode généralisée des Hypervolumes

La méthode généralisée des Hypervolumes suppose que les points $x_1, ..., x_n$ sont générés par une processus de Poisson non homogène N d'intensité $q(\cdot)$ dans $D \subset \mathbb{R}^p$, où D est l'union de k domaines convexes disjoints $D_1, ..., D_k$.

Le problème est alors d'estimer les domaines inconnus D_i dans lesquels les points ont été générés.
Grâce à la propriété conditionnelle du processus de Poisson non homogène, nous pouvons écrire la densité du processus comme :

$$f(x) = \frac{q(x)\, I_D(x)}{\displaystyle\int_D q(t)\, m\,(dt)} = \frac{q(x)\, I_D(x)}{\rho(D)}$$

où $\rho(D) = \displaystyle\int_D q(t)\, m\,(dt)$ est appelé l'intensité intégrée du processus sur D.
La fonction de vraisemblance peut alors s'écrire comme :

$$L_D(x) = \prod_{i=1}^{n} f_X(x_i) = \frac{1}{(\rho(D))^n} \prod_{i=1}^{n} I_D(x_i)\, q(x_i)\ .$$

Le critère généralisé des Hypervolumes est déduit du model statistique, utilisant l'estimation du maximum de vraisemblance.
Soit $\mathcal{P}_k$ l'ensemble des partitions de C en k classes.
Si l'intensité $q(\cdot)$ est connue, la maximisation de la fonction de vraisemblance L_D est équivalente à la minimisation du critère généralisé des Hypervolumes W_k^*.

$$\max_{D_1,\ldots,D_k} L_D(x) \Longleftrightarrow \min_{P\in\mathcal{P}_k} \sum_{i=1}^{k} \rho\left(H(C_i)\right) \Longleftrightarrow \min_{P\in\mathcal{P}_k} W_k^* \ .$$

Le critère généralisé des Hypervolumes est défini par

$$W_k^* = \sum_{i=1}^{k} \rho\left(H(C_i)\right) = \sum_{i=1}^{k} \int_{H(C_i)} q(x)\, m(dx) \ .$$

Nous essayerons donc de trouver la partition P^* telle que

$$P^* = \arg \min_{P\in\mathcal{P}_k} \sum_{l=1}^{k} \int_{H(C_l)} q(x)\, m(dx) \ .$$

4.2.8 Estimation de l'intensité d'un processus de Poisson non homogène

Lorsque l'intensité d'un processus de Poisson n'est pas connue, elle doit être estimée. Pour ce faire nous devons utiliser une méthode non-paramétrique : la méthode des Noyaux.

4.2.9 Passage du processus de Poisson non homogène au processus de Poisson homogène

Par un changement de variables, il est possible de transformer un processus de Poisson non homogène en un processus de Poisson homogène (Cox et Isham (1980)).

Soit $(x_1, ..., x_n)$ une réalisation d'un processus de Poisson non homogène d'intensité $q(x)$.
Nous utilisons le changement de variables suivant :

$$\tau = \tau(x) = \int_0^x q(t)\, m(dt) \text{ sur } \mathbb{R}^p \ .$$

Alors $(\tau_1, ..., \tau_n)$ où $\tau_i = \tau(x_i) \in \mathbb{R}^p$, $(i = 1, ..., n)$ est une réalisation d'un processus de Poisson homogène.

Chapitre 5

Test des Hypervolumes

5.1 Introduction

Dans ce problème de classification, on cherche à déterminer une partition en k classes des données qui minimise le critère des Hypervolumes. Cependant k est fixé, on ne sait pas ce qu'il vaut réellement.

La méthode des Hypervolumes repose, comme on l'a vu, sur un modèle statistique. On a supposé que les points observés résultaient de la réalisation d'un processus de Poisson homogène dans un ensemble D qui est l'union de k domaines convexes disjoints. Ce modèle statistique permet de définir un test du quotient de vraisemblance pour la détermination du nombre de classes.

5.2 Test du quotient de vraisemblance généralisé

5.2.1 Définitions préliminaires

Rappelons tout d'abord la définition du quotient de vraisemblance simple.

Soit $X_1, ..., X_n$ un échantillon aléatoire simple d'une loi $p(x;\theta_0)$ ou $p(x;\theta_1)$. Un test γ de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ est un test du quotient de vraisemblance simple si γ est défini par :

$$\text{on rejette } H_0 \text{ si } \lambda \leq k$$

où

$$\lambda = \lambda(x_1, ..., x_n) = \frac{\prod_{i=1}^{n} p(x_i ; \theta_0)}{\prod_{i=1}^{n} p(x_i ; \theta_1)} = \frac{L(\theta_0; x_1, ..., x_n)}{L(\theta_1; x_1, ..., x_n)} = \frac{L_0}{L_1} \text{ et } k > 0 .$$

La région critique du test s'écrira donc sous la forme

$$W = \left\{ (x_1, ..., x_n) : \frac{L_0}{L_1} \leq k_\alpha \right\} \text{ où } k_\alpha \text{ est tel que } P_{H_0}(W) = \alpha.$$

Rappelons également qu'un test γ^* de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ est appelé un test le plus puissant de niveau α $(0 < \alpha < 1)$ si et seulement si :

- il est de niveau α ;
- $\Pi_{\gamma^*}(\theta_1) \geq \Pi_\gamma(\theta_1)$ quel que soit le test γ pour lequel $\Pi_\gamma(\theta_0) \leq \alpha$.

5.2.2 Lemme de Neyman - Pearson

Soit $X_1, ..., X_n$ un échantillon aléatoire simple d'une loi $p(x; \theta_0)$ ou $p(x; \theta_1)$.
Soit à tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$.
Alors, pour toute valeur de $\alpha \in]0, 1[$, il existe un test le plus puissant de niveau α.
Sa région critique est donnée par

$$W = \left\{ (x_1, ..., x_n) : \lambda = \frac{L(\theta_0; x_1, ..., x_n)}{L(\theta_1; x_1, ..., x_n)} \leq k \right\}$$

où k est une constante strictement positive.

5.2.3 Test du quotient de vraisemblance généralisé

A) Première approche

Soit $X_1, ..., X_n$ un échantillon aléatoire simple d'une loi $p(x; \theta)$.
Un test γ de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est un test du quotient de vraisemblance généralisé si γ est défini par

$$\text{on rejette } H_0 \text{ si } \lambda = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta; x_1, ..., x_n)}{\sup\limits_{\theta \in \Theta} L(\theta; x_1, ..., x_n)} \leq k \, .$$

Et grâce au théorème qui suit, on pourra connaître la région critique du test.

Théorème :

$$\text{Soit } \begin{cases} H_0 : \theta_1 = \theta_1^*; ....; \theta_p = \theta_p^* \ (p \leq n) \\ \\ H_1 = \overline{H}_0 \end{cases}$$

Alors a) La fonction $-2 \ln \lambda$ est asymptotiquement pivotale sous H_0

b) $-2 \ln \lambda \approx \chi^2_p$.

Nous pouvons donc réécrire la région critique comme

$$W = \{(x_1, ..., x_n) : \lambda(x_1, ..., x_n) \leq k\} = \{(x_1, ..., x_n) : -2 \ln \lambda \geq a\} \, .$$

Ce qui implique donc que la région critique du test est connue puisque $a = \chi^2_{p,1-\alpha}$, où $\chi^2_{p,1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi χ^2_p.

Hélas, ce théorème est vrai si l'hypothèse H_0 contient un nombre fini de paramètres ; or, le domaine D à estimer est un paramètre de dimension infinie. Ce théorème ne sera donc pas appliquable.

B) Deuxième approche

Soit $X_1, ..., X_n$ un échantillon aléatoire simple d'une loi $p\,(x; \theta)$.
Un test γ de $H_0 : \ \theta \in \Theta_0$ contre $H_1 : \ \theta \in \Theta_1$ est un test du quotient de vraisemblance généralisé (type II) si γ est défini par

$$\text{on rejette } H_0 \text{ si } \lambda = \frac{\sup\limits_{\theta \in \Theta_1} L(\theta; x_1, ..., x_n)}{\sup\limits_{\theta \in \Theta_0} L(\theta; x_1, ..., x_n)} \geq k \, .$$

5.3 Test des Hypervolumes basé sur le processus de Poisson homogène

Soit $\{x_1, ..., x_n\}$ la réalisation d'un processus de Poisson homogène dans k ensembles convexes disjoints $D_1, ..., D_k$ dans un espace euclidien p-dimensionnel.

Pour $k \geq 2$, nous testons si la subdivision en k classes est significativement meilleure que la subdivision en $k-1$ classes.

Nous testons donc,

H_0 : les données sont générées dans k domaines convexes disjoints,

contre

H_1 : les données sont générées dans $k-1$ domaines convexes disjoints.

Notons par :

- $C = \{C_1, ..., C_k\}$ la partition optimale de $\{x_1, ..., x_n\}$ en k classes.
- $B = \{B_1, ..., B_{k-1}\}$ la partition optimale de $\{x_1, ..., x_n\}$ en $k-1$ classes.

Comme vu précédemment, par la propriété d'uniformité conditionnelle du processus de Poisson, on peut écrire la fonction de vraisemblance pour le domaine D par exemple, comme suit :

$$L(D; x_1, ..., x_n) = \frac{1}{(m(D))^n} I_D\left(H(x_1, ..., x_n)\right) .$$

On utilise la deuxième version du test du quotient de vraisemblance :

$$\lambda(x_1, ..., x_n) = \frac{\sup_B L(B; x_1, ..., x_n)}{\sup_C L(C; x_1, ..., x_n)}$$

ou encore

$$\lambda(x_1,...,x_n) = \frac{\dfrac{1}{\left(\sum_{l=1}^{k-1} m(H(B_l))\right)^n}}{\dfrac{1}{\left(\sum_{l=1}^{k} m(H(C_l))\right)^n}}$$

$$= \left(\frac{W(P,k)}{W(P,k-1)}\right)^n$$

$$= S^n$$

où $S = \dfrac{W(P,k)}{W(P,k-1)}$ et $W(P,k)$ est le critère des Hypervolumes pour une partition de k classes.

La région critique s'écrit sous la forme :

$$W = \{(x_1,...,x_n) : \lambda \geq c_\alpha\} = \{(x_1,...,x_n) : S \geq k_\alpha\} \ .$$

Et on aura comme règle de décision : on rejette H_0, au niveau α, si $S \geq k_\alpha$ où k_α est déterminé en fonction de α.

Auparavant, on ne connaissait malheureusement pas la distribution de la statistique S sous H_0. Nous ne savions donc pas examiner les propriétés théoriques du test ni déterminer le niveau avec lequel on accepte ou rejette l'hypothèse H_0.

Cependant, comme $S \in [0,1[$, car $W(P,k) \leq W(P,k-1)$, en pratique, on rejettait H_0 si S est "proche" de 1 .
En effet, dans ce cas, la valeur du critère des Hypervolumes pour une partition en k classes ne sera pas significativement plus petite que la valeur du critère pour une partition en $k-1$ classes.

Ce test était utilisé de manière séquentielle :
si k_0 était la première valeur de k pour laquelle on rejettait H_0, alors on choisisait $(k_0 - 1)$ comme le nombre de classes "naturelles".

Chapitre 6

Utilisation des tests de permutations pour le test des Hypervolumes

Nous avons vu précédemment que la distribution de la statistique du test des Hypervolumes était inconnue. Dès lors, une bonne façon d'effectuer ce test est de recourir à un test de permutations. Il a fallu, pour ce faire, mettre au point un algorithme, en *Matlab*, qui soit capable d'effectuer des réarrangements aléatoires de données et de calculer, pour chaque permutation, la statistique du test des Hypervolumes.

6.1 Contexte

Tout d'abord, nous allons inverser l'hypothèse nulle et alternative du test des Hypervolumes vu précédemment. La notion d'échangeabilité des données doit être vérifiée sous H_0. C'est le cas si H_0 est l'hypothèse que les données sont générées dans un domaine convexe et non dans deux domaines convexes disjoints.

Le test des Hypervolumes consiste, dès lors, à tester :

H_0 : les données sont générées dans 1 domaine convexe D.
contre
H_1 : les données sont générées dans 2 domaines convexes disjoints, D_1 et D_2.

Appelons C (respectivement C_1, C_2), l'ensemble des points appartenant à D (respectivement D_1 et D_2).

On choisira comme statistique du test :

$$S = \frac{m(H(C_1)) + m(H(C_2))}{m(H(C))}$$

où $m(H(C_1))$, $m(H(C_2))$ et $m(H(C))$ sont respectivement la mesure de Lebesgue de l'enveloppe convexe de C, C_1 et C_2.

La règle de décision sera la suivante : on rejetera H_0 si S est "significativement" petit, c'est-à-dire si la P_{calc} à l'aide du test de permutations est inférieure à α, où α est un niveau de signification fixé par l'utilisateur.

L'algorithme effectuera des permutations entre les deux groupes convexes disjoints et ensuite calculera la statistique S du test. Nous obtiendrons la P_{calc} du test ainsi que la distribution approchée de la statistique S.

Notons que, comme il est rappelé dans le chapitre 2, la méthode des Hypervolumes suppose que les points $x_1, ..., x_n$ sont générés par un processus de Poisson homogène dans l'union de k domaines convexes disjoints. On fera dès lors l'hypothèse qu'il est impossible que trois (ou plus de trois) points se trouvent sur une même droite. Pour cette raison, l'algorithme ne prendra pas ce cas particulier en compte.

Dans le cadre de ce mémoire, nous nous limiterons au test des Hypervolumes sur des données à 2 dimensions. L'extension à des espaces de dimensions supérieures se fera suivant le même principe.

6.2 Input-output de l'algorithme *TestHypervolume.m*

On donne en entrée l'ensemble des points (abscisses et ordonnées) contenu dans un vecteur ainsi que la valeur observée de la statistique du test.

L'algorithme sortira le nombre de s_{per} telles que $S_{per} < s_{obs}$, $S_{per} = s_{obs}$ et $S_{per} > s_{obs}$.

Le programme donne également le temps (en secondes) mis par le programme pour donner les résultats (*temps*). Il enregistre également dans un fichier toutes les données s_{per} à partir desquelles on représentera la distribution approchée de la statistique S grâce au logiciel *SAS*.

6.3 L'algorithme *TestHypervolume.m*

Afin de permuter les 2-groupes convexes disjoints, on aurait pu les trouver tous et les lister. Le but de ce mémoire n'étant pas de créer l'algorithme le plus efficace, mais bien de tester la méthode des tests de permutations, plutôt que d'appliquer cette recherche automatique de tous les 2-groupes convexes disjoints possibles, j'ai opté pour une méthode qui tire au hasard deux 2-groupes à partir des données. Ensuite, l'algorithme teste si ces deux 2-groupes n'ont aucune intersection, ni inclusion entre eux.

Si les 2-groupes ne sont pas disjoints, alors on recherche deux autres 2-groupes jusqu'à ce qu'ils soient disjoints. Dans ce dernier cas, on calcule la statistique S du test. Ce sera le résultat d'une permutation, on recommence cette recherche de 2-groupes convexes disjoints ainsi que le calcul de la statistique S jusqu'à avoir le nombre de permutations désiré.

6.3.1 Permutations aléatoires

Comme on l'a rappelé dans le chapitre 3 de ce mémoire, il est important que les permutations soient **aléatoires** lors d'un test. Pour ce faire, l'algorithme est composé de deux tirages de nombres aléatoires.

- Un premier tirage de nombres aléatoires sert à permuter les données qui se trouvent dans un vecteur.

- Le deuxième nombre aléatoire donne le nombre de points dans le premier 2-groupes. Il permettra de couper le vecteur contenant tous les points des données.

Tant que l'on a pas deux groupes convexes disjoints, on tirera de nouveau ces deux nombres aléatoires, de façon à obtenir deux 2-groupes convexes disjoints de la manière la plus aléatoire possible.

6.3.2 Enveloppes convexes

A partir du premier tirage de nombres aléatoires, on obtient une permutation de l'ensemble des points contenu dans un "grand" vecteur. On coupe ensuite ce "grand" vecteur permuté, au niveau du second nombre aléatoire, en deux nouveaux vecteurs.

On calcule ensuite l'enveloppe convexe de ces 2-groupes. Pour ce faire on utilise le logiciel *Matlab* dont la fonction *convhull* calcule l'enveloppe convexe d'un ensemble d'au moins trois points. On donne en entrée les abscisses et les ordonnées des points, et la fonction donne l'enveloppe convexe et son aire en sortie.

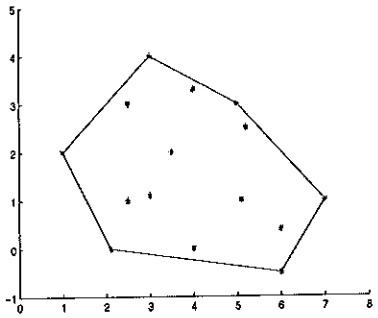


FIG. 6.1 – Fonction *convhull*

6.3.3 Intersection et inclusion

Il faut ensuite vérifier si ces 2-groupes sont disjoints ou non. Pour tester si les enveloppes convexes des deux groupes sont disjoints ou non, on va de nouveau faire appel à des fonctions intégrées dans le logiciel *Matlab*.

La fonction *polybool* est une fonction qui s'occupe uniquement de polygones. On lui donne les coordonnées des sommets des deux polygones ainsi que l'option que l'on veut tester entre ces deux polygones (dans ce cas-ci l'intersection).

Si la fonction nous donne le vide en sortie, c'est que l'aire d'intersection entre les deux ensembles est vide et on calculera la statistique du test pour ces deux groupes. Par contre, si la variable de sortie est différente du vide, c'est que les ensembles ne sont pas disjoints. Ils ne nous intéressent donc pas. On en recherche deux autres.

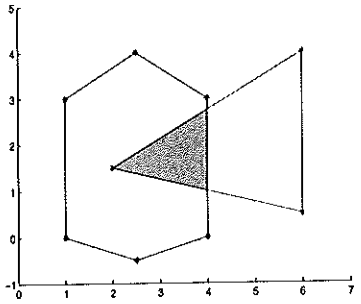


FIG. 6.2 – Fonction *polybool*

Remarques :

Pour pouvoir utiliser la fonction *polybool*, on doit utiliser une autre fonction : (*flatearthpoly*). Celle-ci transforme les coordonnées de l'enveloppe convexe, donnée par la fonction *convhull*, en coordonnées utilisable par *polybool* (projection cylindrique).

Si on n'avait pas eu l'hypothèse que trois points ne peuvent être sur une même droite, il aurait fallu utiliser la fonction *polyxpoly* qui donne les points d'intersection entre deux polygones. En effet la fonction *polybool* ne se préoccupe que de l'aire d'intersection entre deux polygones. Ainsi si deux polygones ne sont joints que par un coté, la fonction *polybool* ne détectera pas d'intersection alors que la fonction *polyxpoly* la détectera.

6.3.4 Cas particuliers

Si on sépare un nuage de points en deux groupes, il se peut qu'un des ensembles ne soit constitué que d'un ou deux points. Pour ces deux cas particuliers, on ne sait pas calculer leur enveloppe convexe à l'aide de la fonction *convhull*. De plus, on ne sait plus utiliser la fonction *polybool* pour tester, dans ce cas-ci, si le point ou la droite intercepte l'autre ensemble de points.

Pour contourner le problème posé par ces deux cas particuliers, il suffit de créer une petite aire autour du point ou de la droite. Ainsi on construit un petit losange, en augmentant ou en soustrayant l'abscisse ou l'ordonnée du point d'un facteur ε très petit.

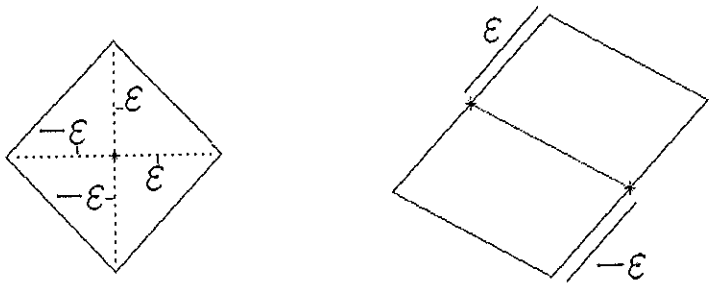


FIG. 6.3 – Aires des cas particuliers

Pour une droite, on augmente et soustrait des deux points situés à l'extrémité de la droite, ε (en abscisse et ordonnée) pour obtenir un mince parallélogramme.

Remarque : on utilise ce petit accroissement de l'aire **uniquement** pour tester si deux groupes sont disjoints ou non. Lorsque l'on calculera la statistique du test, l'aire de l'enveloppe convexe d'un point et d'une droite ainsi modifiée sera remise à zéro.

6.3.5 Temps de calculs

Le temps de calculs pour ce test, (en particulier la sélection de deux groupes disjoints) est bien plus long que pour les tests vus auparavant. C'est pour cette raison que nous nous limiterons à seulement 999 permutations pour ce test.

Chapitre 7

Applications

7.1 Motivation

Nous allons appliquer l'algorithme précédemment vu dans le cas de classification de données. Grâce au test de permutations, nous allons pouvoir valider les deux classes qu'une méthode de classification nous donne. Nous allons arbitrairement choisir la méthode de Ward, comme méthode de classification. Nous exécuterons celle-ci à l'aide du logiciel *SAS*.

Rappelons que l'indice d'agrégation de Ward, consiste à joindre les 2 classes pour lesquelles l'accroissement de l'inertie est minimal.

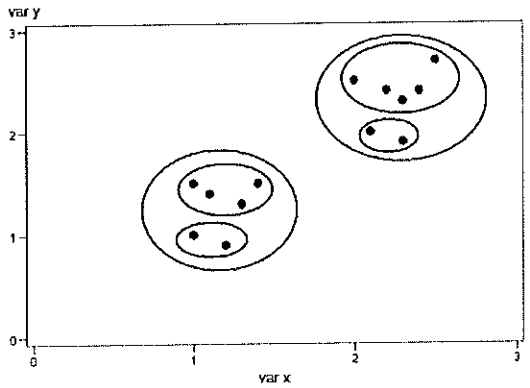


FIG. 7.1 – Représentation de l'indice d'agrégation de Ward

Formellement, l'indice d'agrégation δ est le suivant :

$$\delta(C^l, C^m) = I(C^l \cup C^m) - I(C^l) - I(C^m)$$

où $I(C^l)$ est l'inertie de la classe C^l.

Cet indice d'agrégation favorise la formation de classes sphériques. Il ne donne pas de très bons résultats lorsqu'il traite des données dont la structure est hyperellipsoïdale.

Nous allons traiter trois types de données artificielles dans ce chapitre :

- deux classes hypersphériques
- un ensemble de données sans structure
- deux classes allongées

Pour ces trois applications, nous testerons les hypothèses :

H_0 : les données sont générées dans 1 domaine convexe,

H_1 : les données sont générées dans 2 domaines convexes disjoints,

au niveau de signification $\alpha = 0.05$.

7.2 Première application

7.2.1 Enoncé

Soit un jeu de données contenant deux classes hypersphériques :

X	0.9501	0.2311	0.6068	0.5860	0.8913	0.6621	0.4565	0.0185	0.8214	0.4447
Y	0.6154	0.7919	0.9218	0.7382	0.1763	0.4057	0.9355	0.9169	0.4103	0.6936
X	2.0579	2.3529	2.8132	2.0099	2.1389	2.2028	2.1987	2.6038	2.2722	2.1988
Y	1.0153	1.7468	1.4451	1.9318	1.4660	1.4186	1.8462	1.5252	1.2026	1.6721

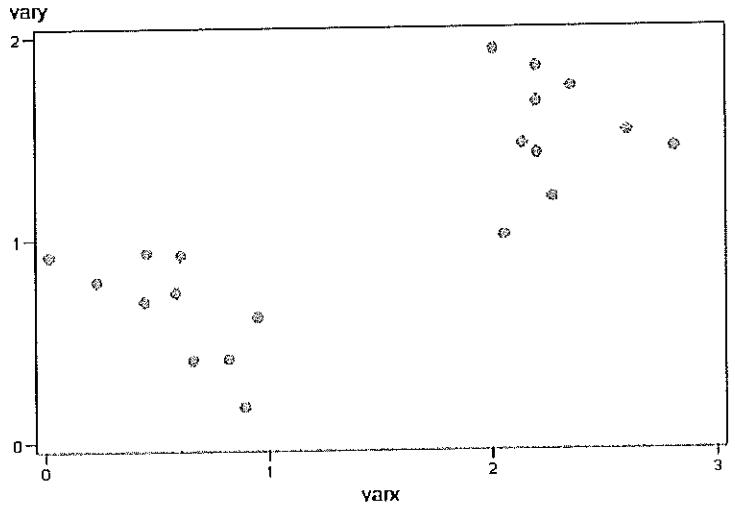


FIG. 7.2 – Nuage des points de la première application

Nous effectuons une analyse de classification avec le logiciel *SAS*.

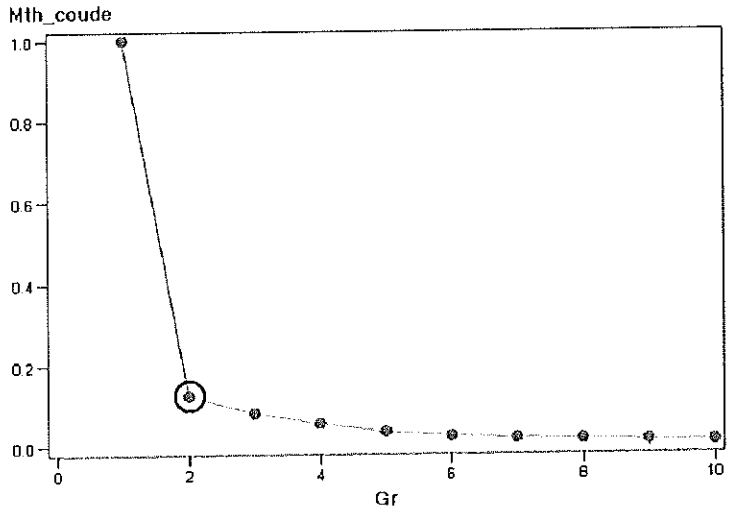


FIG. 7.3 – Graphique de la méthode du coude

La méthode du coude indique clairement le bon nombre de classes à savoir deux.

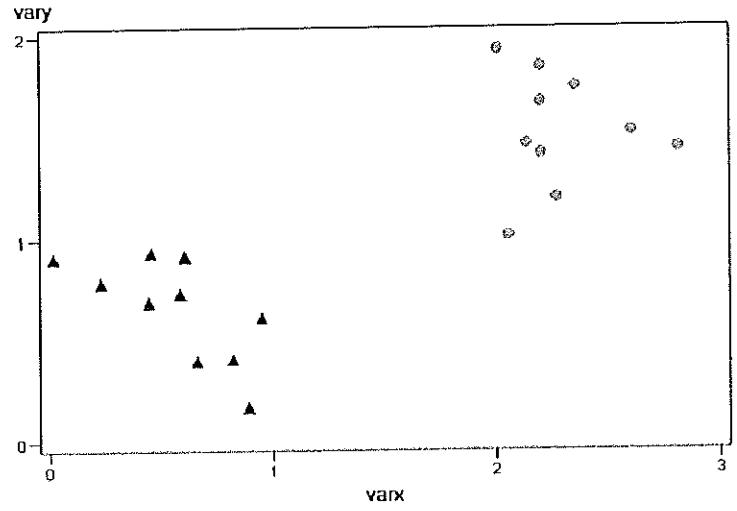


FIG. 7.4 – Classification en 2 classes par la méthode de Ward

De plus, la méthode retrouve les deux classes naturelles.

A partir des deux groupes donnés par la méthode de Ward, nous calculons les aires des enveloppes convexes de ces deux groupes et nous les divisons par l'aire de l'enveloppe convexe de tous les points afin d'obtenir notre statistique observée du test de permutations. Celle-ci vaut : $s_{obs} = 0.3121$.

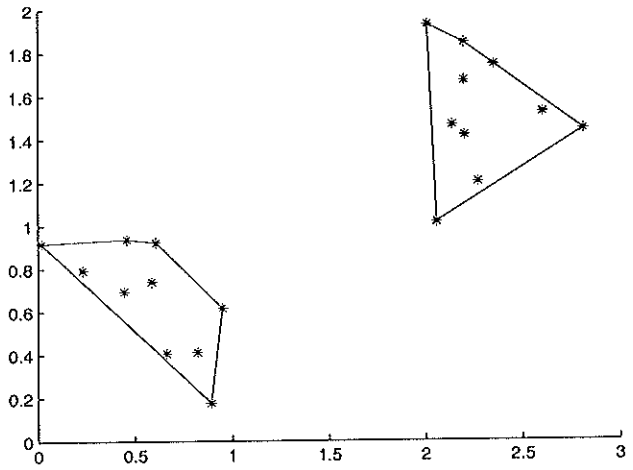
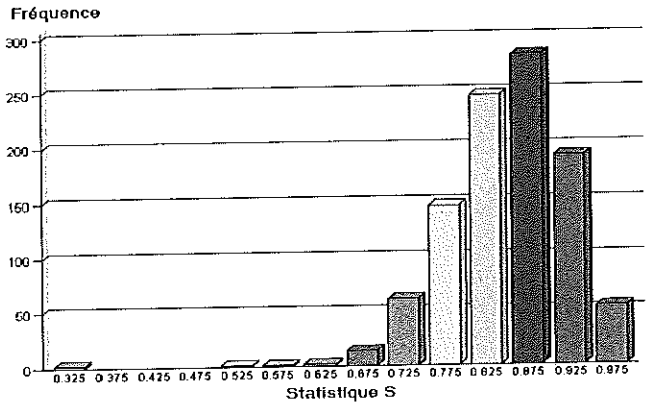


FIG. 7.5 – Enveloppes convexes des deux groupes

7.2.2 test de permutations

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$S_{per} < s_{obs}$	$S_{per} = s_{obs}$	$S_{per} > s_{obs}$
0.3121	3*h*47'21"	0	2	998



$$P_{calc} = \frac{2}{1000} = 0.002 < 0.05 \, .$$

La règle de décision est la suivante :

on rejette H_0 au niveau de signification $\alpha = 0.05$.

7.2.3 Conclusion

Le test confirme, au niveau de signification $\alpha = 0.05$, qu'il existe deux classes dans ce jeu de données.

7.3 Deuxième application

7.3.1 Enoncé

Soit un jeu de données ne contenant sans structure.

X	1.0101	2.0811	1.0021	3.0068	4.0860	5.0913	2.0565	3.0185	4.0214	5.0447
Y	1.0009	1.0129	2.0928	1.0132	1.0099	1.0389	2.0987	2.0038	2.0722	2.0988
X	1.0154	2.0919	3.0218	4.0382	5.0763	1.0057	2.0355	3.0169	4.0103	5.0936
Y	3.0153	3.0468	3.0451	3.0318	3.0660	4.0186	4.0462	4.0252	4.0026	4.0721

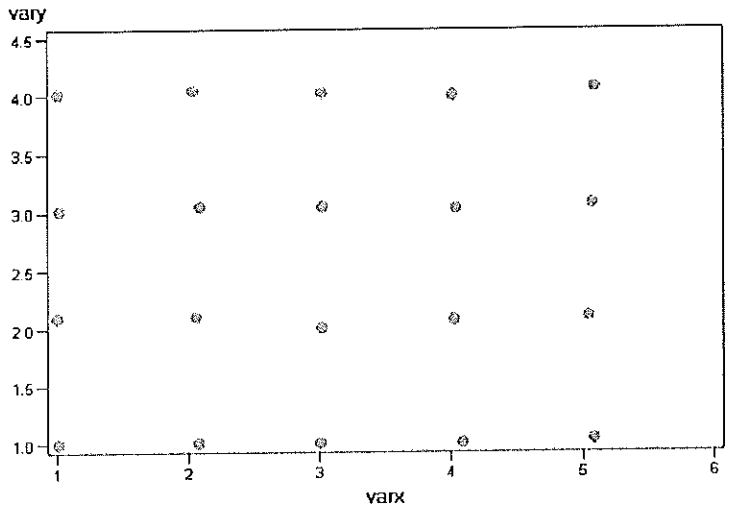


FIG. 7.6 – Nuage des points de la deuxième application

Nous effectuons une analyse de classification avec le logiciel *SAS*.

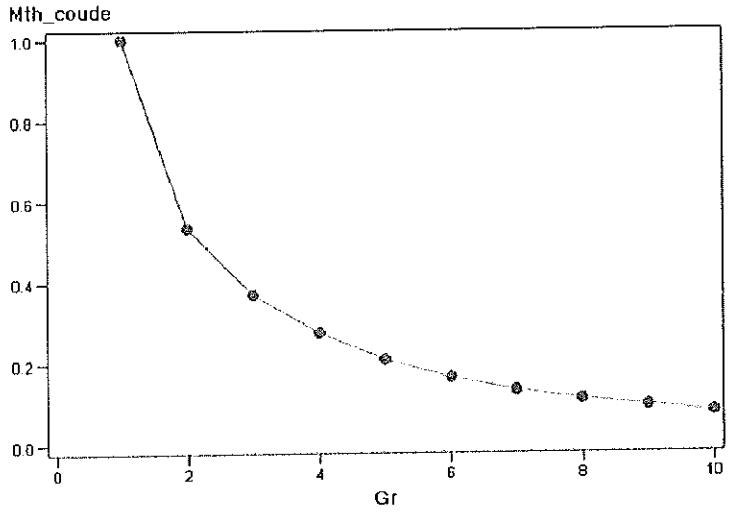


FIG. 7.7 – Graphique de la méthode du coude

La méthode du coude indique qu'il n'y a pas de structure dans les données. Néanmoins, afin d'obtenir la valeur initiale s_{obs} de la statistique utiliser pour le test, nous considérons la classification en deux classes donnée par la méthode de Ward.

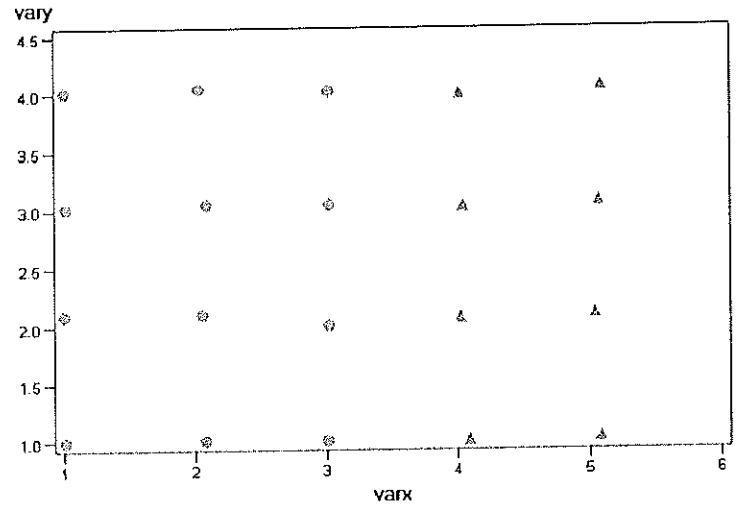


FIG. 7.8 – Classification en 2 classes par la méthode de Ward

Nous obtenons $s_{obs} = 0.7476$.

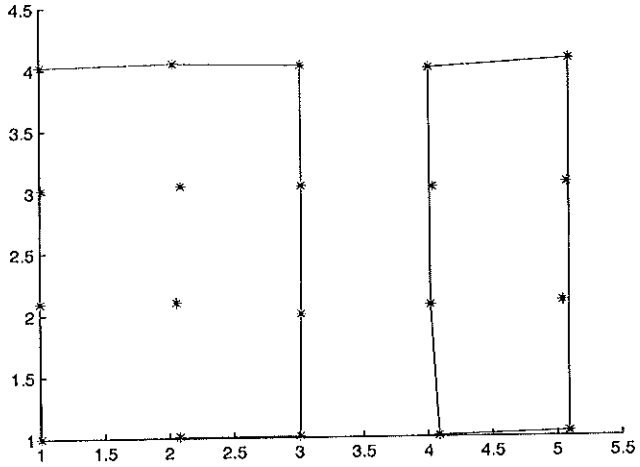
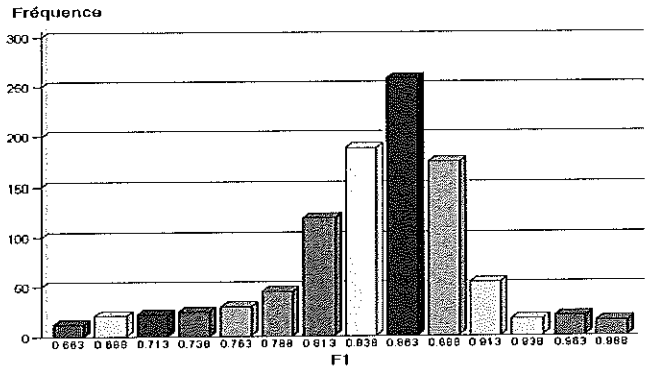


FIG. 7.9 – Enveloppes convexes des deux groupes

7.3.2 test de permutations

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$S_{per} < s_{obs}$	$S_{per} = s_{obs}$	$S_{per} > s_{obs}$
0.7476	$4h06'46"$	80	6	914



$$P_{calc} = \frac{[80] + [6]}{1000} = 0.086 > 0.05$$

La règle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

7.3.3 Conclusion

Le test confirme, au niveau de signification $\alpha = 0.05$, qu'il n'existe pas de structure dans le jeu de données et par conséquent, les deux classes obtenues par la classification de Ward n'ont pas de sens.

7.4 Troisième application

7.4.1 Enoncé

Soit un jeu de données artificielles contenant deux classes allongées.

X	0.56	5.04	2.12	4.61	7.17	6.19	3.78	5.37	3.05	8.10
Y	7.31	7.24	7.01	5.54	7.19	5.34	5.56	5.26	7.17	5.34
X	2.53	4.44	1.68	1.41	8.93	8.59	7.28	5.98	0.56	0.36
Y	5.30	7.11	7.18	5.48	5.53	7.22	5.44	7.01	5.33	5.46

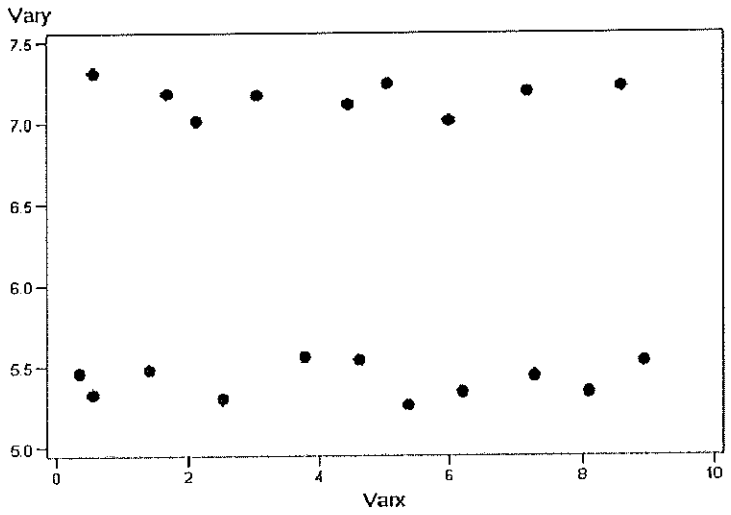


FIG. 7.10 – Nuage des points de la troisième application

Nous effectuons une analyse de classification avec le logiciel *SAS*.

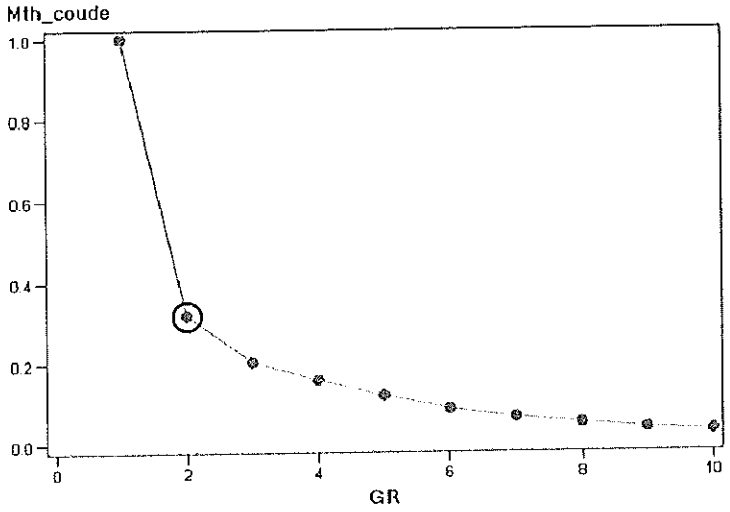


FIG. 7.11 – Graphique de la méthode du coude

La méthode du coude indique le bon nombre de classes, à savoir deux.

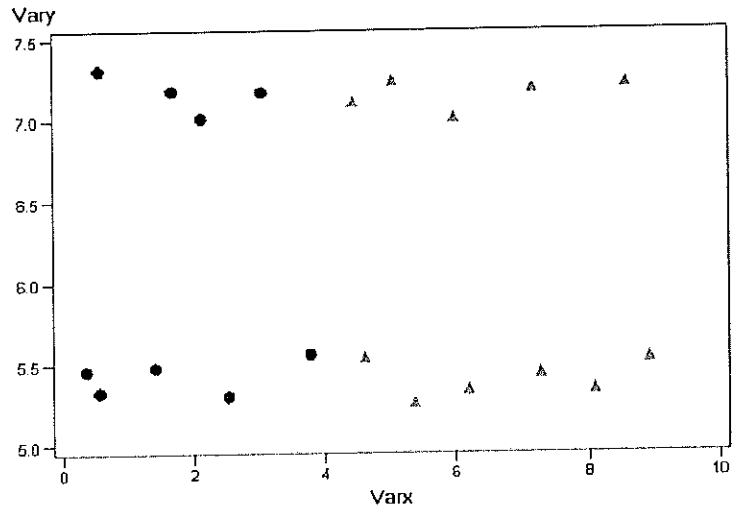


FIG. 7.12 – Classification en 2 classes par la méthode de Ward

Comme on pouvait s'y attendre pour des classes hyperellipsoïdales, la méthode de Ward ne donne pas les deux classes naturelles.

Nous obtenons $s_{obs} = 0.8361$.

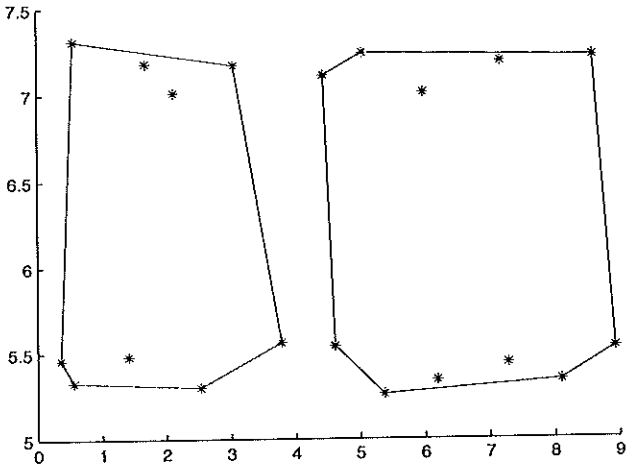
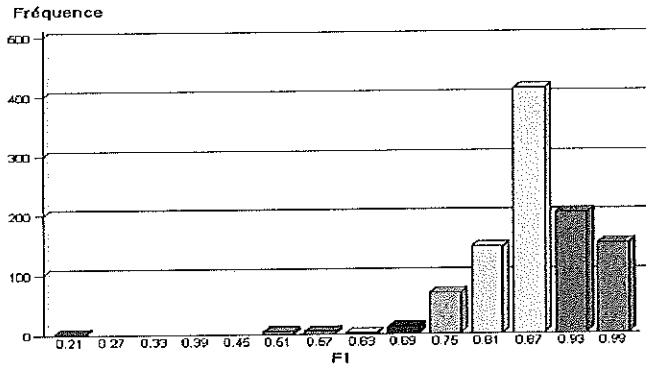


FIG. 7.13 – Enveloppes convexes des deux groupes

7.4.2 test de permutations

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$S_{per} < s_{obs}$	$S_{per} = s_{obs}$	$S_{per} > s_{obs}$
0.8361	$3h54'14"$	233	3	764



$$P_{calc} = \frac{[233] + [3]}{1000} = 0.236 > 0.05$$

La règle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

7.4.3 Conclusion

Ce dernier résultat est très intéressant ; le test refuse la mauvaise classification donnée par Ward, au niveau de signification $\alpha = 0.05$.

Notons que si nous avions pris comme statistique initiale, composée des deux classes naturelles du jeu de données, $s_{obs} = 0.2101$. Dans ce cas là, le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$S_{per} < s_{obs}$	$S_{per} = s_{obs}$	$S_{per} > s_{obs}$
0.8361	$3h54'14"$	0	1	999

On aurait dès lors rejeté H_0 au niveau de signification $\alpha = 0.05$ puisque $P_{calc} = 0.001 < 0.05$.

Ceci tend à montrer que le test des Hypervolumes par test de permutations peut servir à vérifier/valider, si une méthode de classification donne les groupes naturels d'un jeu de données.

Chapitre 8

Détermination du nombre de classes

8.1 Introduction

Dans ce dernier chapitre, nous allons montrer que le test de permutation des Hypervolumes peut aussi être utilisé pour déterminer le nombre de classes d'un jeu de données.

Pour se faire, nous allons utiliser un jeu de données artificielles contenant trois classes distinctes.

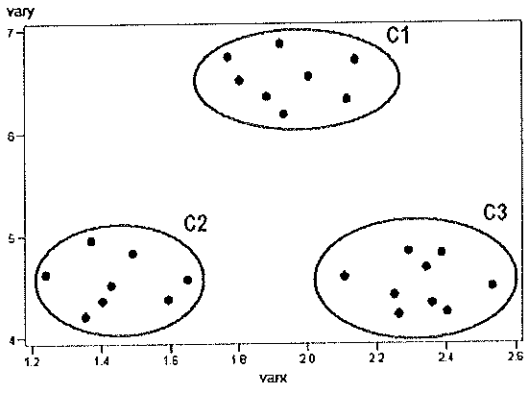


FIG. 8.1 – Données artificielles

Le jeu de données contient 25 points :

X	1.802	1.88	2.003	2.11	1.920	1.93	2.138	1.768	1.65	1.37	1.491	1.24	1.594
Y	6.511	6.35	6.543	6.32	6.864	6.18	6.705	6.737	4.572	4.95	4.832	4.62	4.375

X	1.354	1.428	1.403	2.251	2.107	2.403	2.359	2.342	2.263	2.387	2.533	2.291
Y	4.212	4.518	4.362	4.412	4.593	4.247	4.333	4.678	4.223	4.818	4.491	4.84

Nous allons d'abord séparer l'ensemble des données en deux classes. Nous testerons alors si la division de l'ensemble des données en deux groupes a un sens, à l'aide du test de permutation des Hypervolumes . Comme pour le chapitre précédent, nous obtiendrons s_{obs} en utilisant, arbitrairement, la classification de Ward en deux classes. Nous prendrons comme niveau de signification $\alpha = 0.05$.

Pour chacune des deux classes ainsi obtenues (si H_0 est rejetée), nous allons de nouveau effectuer notre test de permutation des Hypervolumes. Et ainsi de suite pour les groupes ainsi obtenus tant que H_0 n'est pas acceptée.

Si notre méthode fonctionne, on devrait avoir ce schéma-ci :

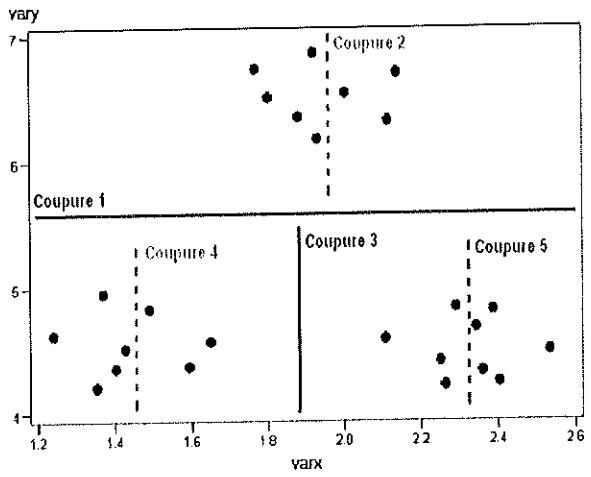


FIG. 8.2 – Schéma des 5 coupures

Où seulement les coupures 1 et 3 seront validées par le test de permutation des Hypervolumes.

8.2 Test pour la première coupure

Nous lançons dans *SAS* une classification avec la méthode de Ward et on lui demande de sortir deux groupes.

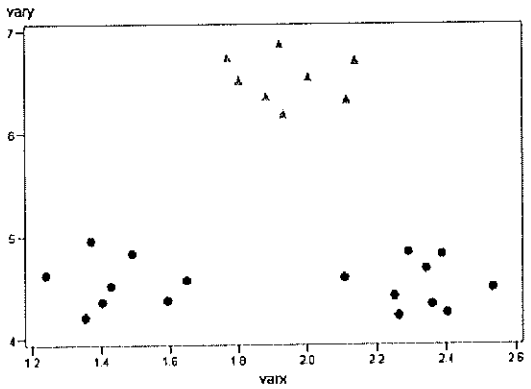


FIG. 8.3 – Classification de Ward pour la première coupure

Nous allons ensuite valider (ou non) cette partition, à partir de cette s_{obs}, avec le programme *TestHypervolume.m.* Il donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$s_{per} < s_{obs}$	$s_{per} = s_{obs}$	$s_{per} > s_{obs}$
0.4250	$4h26'47"$	0	1	999

$$P_{calc} = \frac{1}{1000} = 0.001 < 0.05$$

La règle de décision est la suivante :

on rejette H_0 au niveau de signification $\alpha = 0.05$.

Le test confirme, au niveau de signification $\alpha = 0.05$, la coupure en deux classes de ce jeu de données.

8.3 Test pour la deuxième coupure

La classification de Ward pour la deuxième coupure nous donne :

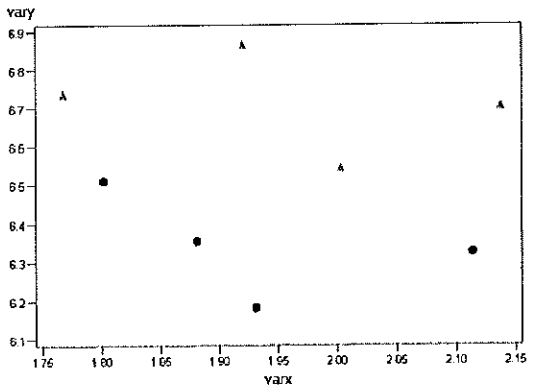


FIG. 8.4 – Classification de Ward pour la deuxième coupure

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$s_{per} < s_{obs}$	$s_{per} = s_{obs}$	$s_{per} > s_{obs}$
0.5799	2'52"	107	13	880

$$P_{calc} = \frac{[107] + [13]}{1000} = 0.12 > 0.05$$

La règle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

Le test confirme, au niveau de signification $\alpha = 0.05$, que la coupure en deux classes de ce sous-groupe n'a pas lieu d'être.

8.4 Test pour la troisième coupure

La classification de Ward pour la troisième coupure nous donne :

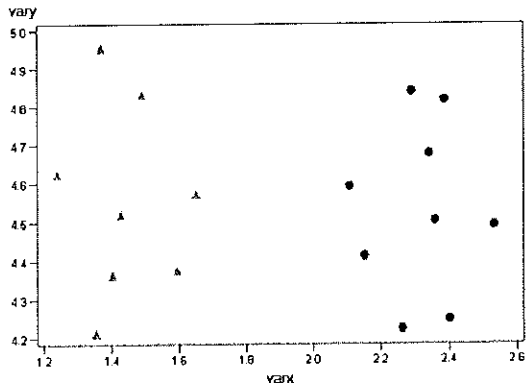


FIG. 8.5 – Classification de Ward pour la troisième coupure

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$s_{per} < s_{obs}$	$s_{per} = s_{obs}$	$s_{per} > s_{obs}$
0.4450	15'22"	0	3	997

$$P_{calc} = \frac{3}{1000} = 0.003 < 0.05$$

La règle de décision est la suivante :

on rejette H_0 au niveau de signification $\alpha = 0.05$.

Le test confirme, au niveau de signification $\alpha = 0.05$, la coupure en deux classes du sous-groupe.

8.5 Test pour la quatrième coupure

La classification de Ward pour la quatrième coupure nous donne :

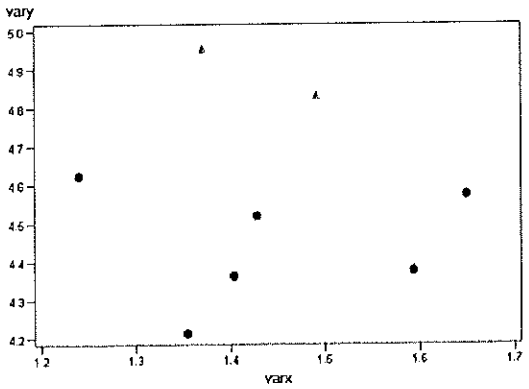


FIG. 8.6 – Classification de Ward pour la quatrième coupure

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$s_{per} < s_{obs}$	$s_{per} = s_{obs}$	$s_{per} > s_{obs}$
0.5661	2'15"	156	28	816

$$P_{calc} = \frac{[156] + [28]}{1000} = 0.184 > 0.05$$

La règle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

Le test confirme, au niveau de signification $\alpha = 0.05$, que la coupure en deux classes de ce sous sous-groupe n'a pas lieu d'être.

8.6 Test pour la cinquième coupure

La classification de Ward pour la cinquième coupure nous donne :

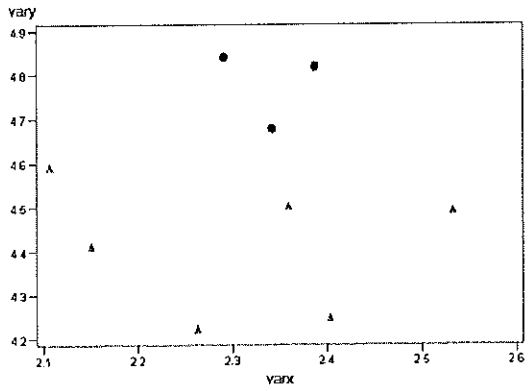


FIG. 8.7 – Classification de Ward pour la cinquième coupure

Le programme *TestHypervolume.m* donne, pour 999 permutations aléatoires, les résultats suivants :

s_{obs}	*temps*	$s_{per} < s_{obs}$	$s_{per} = s_{obs}$	$s_{per} > s_{obs}$
0.5916	2'09"	80	15	905

$$P_{calc} = \frac{[80] + [15]}{1000} = 0.095 > 0.05$$

La règle de décision est la suivante :

on ne rejette pas H_0 au niveau de signification $\alpha = 0.05$.

Le test confirme, au niveau de signification $\alpha = 0.05$, que la coupure en deux classes de ce sous sous-groupe n'a pas lieu d'être.

8.7 Conclusion

A partir d'une méthode de classification qui détermine la statistique initiale du test s_{obs}, on retrouve les trois classes naturelles de notre jeu de données.

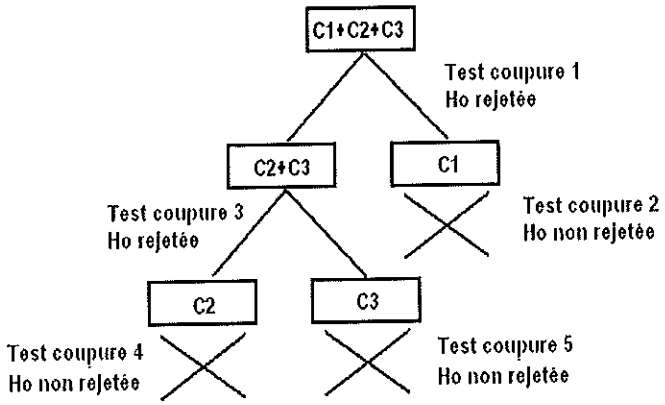


FIG. 8.8 – Arbre de classification

Nous avons choisi arbitrairement la méthode de Ward, mais les autres méthodes telles que, par exemple, du lien simple, des centroïdes, auraient donné les mêmes résultats.

Nous venons de voir que le test de permutations des Hypervolumes permettait de déterminer le bon nombre de classes dans le cadre d'une procédure hiérarchique divisive.

Conclusion

L'objectif de ce mémoire était, dans un premier temps, de comparer les tests d'hypothèses dit classiques et les tests de permutations. Pour ce faire, nous les avons appliqués à quatre exemples. Afin de pouvoir utiliser les tests de permutations, il a fallu implémenter, pour chaque test d'hypothèse, un algorithme. Ceux-ci ont été réalisés grâce au logiciel *Matlab*.

Les conclusions obtenues, pour les exemples pris dans ce travail, que ce soit pour les tests classiques ou de permutations, sont les mêmes. Pour un exemple donné, les deux tests conduisent aux mêmes décisions d'acceptation ou de rejet de l'hypothèse nulle, au niveau de signification α.

Pour les quatre exemples, on remarquera également que l'histogramme des valeurs de la statistique permutée approche valablement la fonction de distribution théorique de la statistique du test classique.

La seconde partie du mémoire traitait de classification et plus précisément du test des Hypervolumes.

Avant ce mémoire, le test des Hypervolumes était utilisé de manière heuristique, car on ne connaissait pas la distribution de la statistique du test. Pour combler ce vide, nous avons effectué le test des Hypervolumes à l'aide d'un test de permutations. Un programme a été spécialement conçu en *Matlab* à cet effet et, grâce à lui, nous avons pu calculer des p-valeurs (approchées) pour la statistique du test des Hypervolumes.

D'autre part, un problème souvent rencontré en classification vient du fait que certaines méthodes de classification sont biaisées. Par exemple, la méthode de Ward ne trouve pas les classes naturelles lorsqu'il est utilisé sur des données dont la structure est constituée classes allongées.

Il est dès lors très intéressant de pouvoir tester si une méthode de classification donne les bons groupes. Et nous avons justement vu, dans le septième chapitre, que le test de permutations des Hypervolumes permet de valider la division d'une classe en deux sous-classes.

Bien plus, le dernier chapitre montre que le test de permutations des Hypervolumes permet également de déterminer le nombre de classes contenues dans un jeu de données, lorsque ces classes sont obtenues par une méthode hiérarchique divisive de classification.

Rappelons, toutefois, que le test des Hypervolumes a pour hypothèse très contraignante le fait que les deux groupes, que l'on teste, doivent être convexes. Cette limite ne nous permettra donc pas de traiter tous les types de données avec cette méthode.

Notons qu'à l'avenir, il serait très intéressant de généraliser le test de permutations des Hypervolumes à 3 puis à p dimensions.

Il faudrait également améliorer l'algorithme *TestHypervolume.m*. Le but de ce mémoire n'était pas de construire l'algorithme le plus efficace mais bien, dans un premier temps, de vérifier si ce test donnait des résultats intéressants. Cela étant fait, il faudrait maintenant le perfectionner, afin de pouvoir traiter de plus gros ensembles de données, en un temps acceptable. Une façon de procéder, serait peut-être, non plus de tester aléatoirement si deux groupes sont disjoints ou non, mais plutôt, de lister tous les 2-groupes convexes disjoints, avant de faire le test de permutations.

Annexe A

Programmes utilisés

A.1 Testdifmoyenneconnue.m

```
1  function y = Testdifmoyenneconnue(X,Y,var1,var2)
2  % === fonction qui permute 2 échantillons ====
3
4  %==input==
5  % X represente le 1er échantillon, Y le 2eme
6  % var1 et var2 la variance de la population 1 et population 2.
7  %==output==
8  % sobs; la valeur observée de la stat
9  % le fichier "résultat.doc" contenant tous les résultats des statistiques
10 % après permutations
11 % temps : le temps mis par l'ordinateur pour fournir tous les résultats
12 % Résultats du test qui indique le nombre de valeurs comprises en desous de
13 % la valeur -|sobs| : premier, égal à -|sobs|: deuxieme, comprise entre
14 % -|sobs| et |sobs|:troisieme, égal à |sobs|: quatrieme
15 %et enfin plus grande que |sobs|:cinquieme
16
17
18 Doc=fopen('Résultat.doc','w');%nom du fichier dans lequel se trouve les résultats.
19 nbela=length(X) ; % nbela le nombre d'éléments de l'échantillon X et nbelb de Y.
20 nbelb =length(Y);
21 nbel= nbela + nbelb;
22
23 % Utilisateur choisit le nombre de permutations aléatoires souhaitées: nbp
24 nbp = input ('donnez le nombre de permutations aléatoires voulues: ');
25 disp ('echantillon 1:')
26 X % affiche l'échantillon X
```

```
disp ('echantillon 2:')
Y % affiche l'échantillon Y

%Initialisation des variables
S=0;
T=0;
premier = 0;
deuxieme = 0;
troisieme = 0;
quatrieme = 1;  % car on ajoute la valeur initiale
cinquieme = 0;
for i= 1:nbela
    S=S+X(i);
end
for i= 1:nbelb
       T=T+Y(i);
end
Moyenneinit1 = S/nbela;
Moyenneinit2 = T/nbelb;
statinit = Moyenneinit1 -Moyenneinit2;
% Statistique observée:
zobs = statinit/sqrt((var1/nbela)+(var2/nbelb))
fprintf(Doc,'%8.4f',zobs);%indique le résultat dans "résultat.doc"

%===programme qui permute aléatoirement les éléments de 2 tableaux===
Z = [X Y];
for j=1:nbp  %boucle des permutations
R = randperm(nbel); % Donne une permutation des nombres de 1 jusque nbel
S1=0;
S2=0;
for i=1:nbel   % Echange les éléments des 2 tableaux
    P(i) = Z(R(i));
end
for i= 1:nbela
  A(i)=P(i);
  S1 = S1+A(i);
end
for i = nbela+1:nbel
  B(i-nbela)=P(i);
   S2 = S2+ P(i);
end

j;
```

```
70  A;
71  B;
72  Moyenne1 = S1/nbela;
73  Moyenne2 = S2/nbelb;
74  statper = Moyenne1 - Moyenne2;
75  % Statistique observée après permutations
76  Sper = statper/sqrt((var1/nbela)+(var2/nbelb));
77  fprintf(Doc,'\n');
78  fprintf(Doc,'%8.4f',Sper); %indique le résultat dans "résultat.doc"
79
80  % Test qui fournit le nombre de tper se trouvant dans premier deuxieme troisieme
81  % quatrieme cinquieme
82  if Sper < - abs(zobs)
83      premier = premier + 1;
84  elseif Sper == - abs(zobs)
85     deuxieme = deuxieme + 1;
86  elseif Sper ==  abs(zobs)
87     quatrieme = quatrieme + 1;
88  elseif Sper >  abs(zobs)
89    cinquieme =  cinquieme + 1;
90  else
91      troisieme = troisieme +1;
92  end
93
94  end %fin boucle des permutations
95
96  temps=cputime
97
98  premier
99  deuxieme
100 troisieme
101 quatrieme
102 cinquieme
103
104 fclose(Doc);
```

A.2 Testdifmoyenneinconnue.m

```
1   function y = Testdifmoyenneinconnue(X,Y)
2   % === fonction qui permute 2 échantillons ====
3
4   %==input==
5   % X represente le 1er échantillon, Y le 2eme
6
7   %==output==
8   % tobs; la valeur observée de la stat
9   % le fichier "résultat.doc" contenant tous les résultats des statistiques
10  % après permutations
11  % temps : le temps mis par l'ordinateur pour fournir tous les résultats
12  % Résultats du test qui indique le nombre de valeurs comprises en desous de
13  % la valeur -|tobs| : premier, égal à -|tobs|: deuxieme, comprise entre
14  % -|tobs| et |tobs|:troisieme, égal à |tobs|: quatrieme
15  %et enfin plus grande que |tobs|:cinquieme
16
17
18  Doc=fopen('Résultat.doc','w');%nom du fichier dans lequel se trouve les résultats.
19  nbela=length(X) ;% nbela le nombre d'éléments de l'échantillon X et nbelb de Y.
20  nbelb =length(Y);
21  nbel= nbela + nbelb;
22
23  % Utilisateur choisit le nombre de permutations aléatoires souhaitées: nbp
24  nbp = input ('donnez le nombre de permutations aléatoires voulues: ');
25  disp ('echantillon 1:')
26  X  % affiche l'échantillon X
27  disp ('echantillon 2:')
28  Y % affiche l'échantillon Y
29
30  %Initialisation des variables
31  S=0;
32  T=0;
33  premier = 0;
34  deuxieme = 0;
35  troisieme = 0;
36  quatrieme = 1;  % car on ajoute la valeur initiale
37  cinquieme = 0;
38  for i= 1:nbela
39      S=S+X(i);
```

```
40  end
41  for i= 1:nbelb
42        T=T+Y(i);
43  end
44  Moyenneinit1 = S/nbela;
45  Moyenneinit2 = T/nbelb;
46  statinit = Moyenneinit1 -Moyenneinit2;
47  Spd= ((nbela-1)*var(X)+(nbelb-1)*var(Y))/(nbel-2);
48  % Statistique observée:
49  tobs = statinit/ sqrt(Spd)
50  fprintf(Doc,'%8.4f',tobs);%indique le résultat dans "résultat.doc"
51
52  %===programme qui permute aléatoirement les éléments de 2 tableaux===
53  Z = [X Y];
54  for j=1:nbp  %boucle des permutations
55  R = randperm(nbel);
56  S1=0;
57  S2=0;
58  for i=1:nbel
59      P(i) = Z(R(i));
60  end
61  for i= 1:nbela
62    A(i)=P(i);
63    S1 = S1+A(i);
64  end
65  for i = nbela+1:nbel
66    B(i-nbela)=P(i);
67     S2 = S2+ P(i);
68  end
69
70  j;
71  A;
72  B;
73  Moyenne1 = S1/nbela;
74  Moyenne2 = S2/nbelb;
75  statper = Moyenne1 - Moyenne2;
76  % Statistique observée après permutations
77  Spdper=  ((nbela-1)*var(A)+(nbelb-1)*var(B))/(nbel-2);
78  tper=statper/ sqrt(Spdper);
79  fprintf(Doc,'\n');
80  fprintf(Doc,'%8.4f',tper);  %indique le résultat dans "résultat.doc"
81
82  a= abs(tper-tobs);
```

```
83  b= abs(tper+tobs);
84
85  % Test qui fournit le nombre de tper se trouvant dans premier deuxieme troisieme
86  % quatrieme cinquieme
87  if tper < - abs(tobs)
88      premier = premier + 1;
89  elseif tper == - abs(tobs)
90     deuxieme = deuxieme + 1;
91  elseif b <= 0.0001 % changement de type de condition du aux erreurs d'arrondis
92       deuxieme = deuxieme + 1;
93  elseif tper ==  abs(tobs)
94     quatrieme = quatrieme + 1;
95  elseif a <= 0.0001 % changement de type de condition du aux erreurs d'arrondis
96        quatrieme = quatrieme + 1;
97  elseif tper >  abs(tobs)
98        cinquieme =  cinquieme + 1;
99  else
100     troisieme = troisieme +1;
101 end
102
103 end  %fin boucle des permutations
104
105 temps=cputime
106
107 premier
108 deuxieme
109 troisieme
110 quatrieme
111 cinquieme
112
113 fclose(Doc);
```

A.3 Testrapvariance.m

```
1  function y = Testrapvariance(X,Y)
2  % === fonction qui permute 2 échantillons ====
3
4  format long
5  %==input==
6  % X represente le 1er échantillon, Y le 2eme
7
8  %==output==
9  % sobs; la valeur observée de la stat
10 % le fichier "résultat.doc" contenant tous les résultats des statistiques
11 % après permutations
12 % temps : le temps mis par l'ordinateur pour fournir tous les résultats
13 % Résultats du test qui indique le nombre de valeurs comprises en desous de
14 % la valeur -|sobs| : premier, égal à -|sobs|: deuxieme, comprise entre
15 % -|sobs| et |sobs|:troisieme, égal à |sobs|: quatrieme
16 %et enfin plus grande que |sobs|:cinquieme
17
18 Doc=fopen('Résultat.doc','w');%nom du fichier dans lequel se trouve les résultats.
19 nbela=length(X) ;% nbela le nombre d'éléments de l'échantillon X et nbelb de Y.
20 nbelb =length(Y);
21 nbel= nbela + nbelb;
22
23 % Utilisateur choisit le nombre de permutations aléatoires souhaitées: nbp
24 nbp = input ('donnez le nombre de permutations aléatoires voulues: ');
25 disp ('echantillon 1:')
26 X
27 disp ('echantillon 2:')
28 Y
29
30 % Initialisation des variables
31 premier = 0;
32 deuxieme = 0;
33 troisieme = 0;
34 quatrieme = 1;  % car on ajoute la valeur initiale
35 cinquieme = 0;
36 % Statistique observée:
37 statinit= ((nbela-1)*var(X))/((nbelb-1)*var(Y))
38 fprintf(Doc,'%8.4f',statinit);%indique le résultat dans "résultat.doc"
39
```

```
40
41  %===programme qui permute aléatoirement les éléments de 2 tableaux===
42  Z = [X Y];
43  for j=1:nbp  %boucle des permutations
44  R = randperm(nbel);% Donne une permutation des nombres de 1 jusque nbel
45
46  for i=1:nbel  % Echange les éléments des 2 tableaux
47      P(i) = Z(R(i));
48  end
49  for i= 1:nbela
50    A(i)=P(i);
51  end
52  for i = nbela+1:nbel
53    B(i-nbela)=P(i);
54  end
55
56  j;
57  A;
58  B;
59  % Statistique observée après permutations
60  statper=((nbela-1)*var(A))/((nbelb-1)*var(B))  ;
61  fprintf(Doc,'\n');
62  fprintf(Doc,'%8.4f',statper);  %indique le résultat dans "résultat.doc"
63
64  %statper = 0.87993010
65  a= abs(statper-statinit);
66
67
68  % Test qui fournit le nombre de tper se trouvant dans premier deuxieme troisieme
69  % quatrieme cinquieme
70  if statper < - abs(statinit)
71      premier = premier + 1;
72  elseif statper == - abs(statinit)
73     deuxieme = deuxieme + 1;
74  %elseif statper == statinit
75   %      quatrieme = quatrieme + 1;
76  elseif a <= 0.001 % changement de type de condition du aux erreurs d'arrondis
77        quatrieme = quatrieme + 1;
78  elseif statper >  abs(statinit)
79    cinquieme =  cinquieme + 1;
80  else
81      troisieme = troisieme +1;
82  end
```

```
83
84  end %fin boucle des permutations
85
86  temps=cputime
87
88  premier
89  deuxieme
90  troisieme
91  quatrieme
92  cinquieme
93
94  fclose(Doc);
```

A.4 Testcorrelation.m

```
1  function y = Testcorrelation(X,Y)
2  % === fonction qui permute 2 échantillons ====
3
4  %==input==
5  % X represente les abscisses de 1 échantillon, Y les ordonnées
6
7  %==output==
8  % tobs; la valeur observée de la stat
9  % le fichier "résultat.doc" contenant tous les résultats des statistiques
10 % après permutations
11 % temps : le temps mis par l'ordinateur pour fournir tous les résultats
12 % Résultats du test qui indique le nombre de valeurs comprises en desous de
13 % la valeur -|tobs| : premier, égal à -|tobs|: deuxieme, comprise entre
14 % -|tobs| et |tobs|:troisieme, égal à |tobs|: quatrieme
15 %et enfin plus grande que |tobs|:cinquieme
16
17 Doc=fopen('Résultat.doc','w');%nom du fichier dans lequel se trouve les résultats.
18 nbel=length(X) ; %nbel le nombre d'éléments de l'échantillon
19
20 % Utilisateur choisit le nombre de permutations aléatoires souhaitées: nbp
21 nbp = input ('donnez le nombre de permutations aléatoires voulues: ');
22 disp ('abscisse de 1 echantillon :')
23 X
24 disp ('ordonnée de 1 echantillon :')
25 Y
26 premier = 0;
27 deuxieme = 0;
28 troisieme = 0;
29 quatrieme = 1;  % car on ajoute la valeur initiale
30 cinquieme = 0;
31 Minit1 = mean(X);
32 Minit2 = mean(Y);
33 F=0;
34 G=0;
35 H=0;
36 for i = 1:nbel
37 F = F + ( (X(i)-Minit1)* (Y(i)-Minit2) );
38 G = G + ( X(i) - Minit1)^2;
39 H  = H + ( Y(i) - Minit2)^2;
```

```
40  end
41  robs = F / ( sqrt(G)* sqrt(H));
42  % Statistique observée:
43  statinit = sqrt(nbel - 2)* robs / sqrt(1- robs^2)
44  fprintf(Doc,'%8.4f',statinit); %indique le résultat dans "résultat.doc"
45
46  %===programme qui permute aléatoirement les éléments de 2 tableaux===
47
48  for j=1:nbp  %boucle des permutations
49  R = randperm(nbel);% Donne une permutation des nombres de 1 jusque nbel
50  A=X(R); % permutation de X
51  B= Y; % Pour ce type de test Y reste fixé
52  M1 = mean(A);
53  M2 = mean(B);
54  I=0;
55  J=0;
56  K=0;
57  for i = 1:nbel
58  I = I + ( (A(i)-M1) *(B(i)-M2) );
59  J = J + ( A(i) - M1)^2;
60  K  = K + ( B(i) - M2)^2 ;
61  end
62  rper = I / ( sqrt(J)* sqrt(K));
63  % Statistique observée après permutations:
64  statper = sqrt(nbel - 2)* rper / sqrt(1- rper^2);
65  fprintf(Doc,'\n');
66  fprintf(Doc,'%8.4f',statper);  %indique le résultat dans "résultat.doc"
67
68  % Test qui fournit le nombre de tper se trouvant dans premier deuxieme troisieme
69  % quatrieme cinquieme
70  if statper < - abs(statinit)
71      premier = premier + 1;
72  elseif statper == - abs(statinit)
73     deuxieme = deuxieme + 1;
74  elseif statper ==  abs(statinit)
75     quatrieme = quatrieme + 1;
76  elseif statper >  abs(statinit)
77    cinquieme =  cinquieme + 1;
78  else
79      troisieme = troisieme +1;
80  end
81
82  end %fin boucle des permutations
```

```
83
84   temps=cputime
85
86   premier
87   deuxieme
88   troisieme
89   quatrieme
90   cinquieme
91
92   fclose(Doc);
```

A.5 TestHypervolume.m

```
1  function y = TestHypervolume(X,statinit)
2  % Test de permutations des Hypervolumes
3
4  %==input==
5  % X est une matrice 2 x nbel qui contient les abscisses et les ordonnées de
6  % l'ensemble des points du jeu de données.
7  % statinit est la valeur de la statistique observée déterminée par une
8  % méthode de classification.
9
10 %==output==
11 % sobs = statinit
12 % le fichier "résultat.doc" contenant tous les résultats des statistiques
13 % après permutations
14 % temps : le temps mis par l'ordinateur pour fournir tous les résultats
15 % Résultats du test qui indique le nombre de valeurs permutées sper
16 % telles que sper < sobs, sper = sobs et sper > sobs.
17
18 format long
19 Doc=fopen('Résultat.doc','w');%nom du fichier dans lequel se trouve les résultats.
20 fprintf(Doc,'%2.8f',statinit);  %indique le résultat dans "résultat.doc"
21
22 %%%%% Constantes %%%%%
23 nbel=length(X);   % nombres d'éléments
24 bol=0;    % Booléan
25 epsilon=0.001;  % petit accroisement d'aire
26 premier = 0;
27 deuxieme = 0;
28 troisieme = 0;
29 quatrieme = 1;  % car on ajoute la valeur initiale
30 cinquieme = 0;
31
32 nbp = input ('donnez le nombre de permutations aléatoires voulues: ');
33
34 [xdebut,airetot] = convhull(X(:,1),X(:,2));
35 airetot  % Donne l'aire de l'enveloppe convexe de l'ensemble des points
36
37 for permu=1:nbp     %boucle nbre de permu
38
39
```

```
40  while bol == 0      %boucle: tant que pas 2-groupes convexes disjoints
41
42  R = randperm(nbel); %permutation des differents points
43
44  for i= 1:nbel
45  Y(i,:)= X(R(i),:);
46  end
47
48  P = randperm(nbel-1); % permutation du nombre de pts dans les ensembles
49
50  for j = 1 : P(1)      % Séparation du "grand" vecteur en 2
51   A(j,:)= Y(j,:);
52  end
53
54  for e= P(1)+1: nbel
55     B((e-P(1)),:) = Y(e,:);
56  end
57
58
59  if P(1)== 1             % ensemble = 1pt
60      aire=1;
61      C(1,:)=[(A(1,1)+epsilon) A(1,2)]; %On "élargit" un peu le point
62      C(2,:)=[A(1,1) (A(1,2)+epsilon)];
63      C(3,:)=[(A(1,1)-epsilon) A(1,2)];
64      C(4,:)=[A(1,1) (A(1,2)-epsilon)];
65      A=C;
66
67  elseif P(1) ==2         % ensemble = une droite
68      aire=1;
69      C(1,:)=A(1,:)+ epsilon ;  % On va augmenter la droite et creer un rectangle
70      C(2,:)=A(2,:)+ epsilon;
71      C(3,:)=A(1,:)- epsilon;
72      C(4,:)=A(2,:)- epsilon ;
73      A=C;
74
75  elseif P(1) ==(nbel-2)        % ensemble = une droite
76      aire=2;
77      C(1,:)=B(1,:)+ epsilon;   % On va augmenter la droite et creer un rectangle
78      C(2,:)=B(2,:)+ epsilon;
79      C(3,:)=B(1,:)- epsilon;
80      C(4,:)=B(2,:)- epsilon ;
81      B=C;
82
```

```
83  elseif P(1)==(nbel-1)           % ensemble = 1pt
84      aire=2;
85      C(1,:)=[(B(1,1)+epsilon) B(1,2)]; %On "élargit" un peu le point
86      C(2,:)=[B(1,1) (B(1,2)+epsilon)];
87      C(3,:)=[(B(1,1)-epsilon) B(1,2)];
88      C(4,:)=[B(1,1) (B(1,2)-epsilon)];
89      B=C;
90
91  else                    % 2 ensembles où on sait calculer l'aire
92      aire=3;
93
94  end %if
95
96
97  hold on
98  %axis([-1 5 0 8]);
99
100 a1=A(:,1);
101 a2=A(:,2);
102 k = convhull(a1,a2);  % donne l'aire et les pts composant l'enveloppe convexe de A
103 %plot(a1(k),a2(k),'r-',a1,a2,'r*')
104
105
106 b1=B(:,1);
107 b2=B(:,2);
108 h = convhull(b1,b2) ;  % donne l'aire et les pts composant l'enveloppe convexe de
109 %plot(b1(h),b2(h),'b-',b1,b2,'b*')
110
111 A=[ ];    % On vide A et B
112 B=[ ];
113
114 [inty,intx] = polyxpoly(a2(k),a1(k),b2(h),b1(h));
115 %donne les intersections entre les 2 enveloppes convexes
116
117 [lat1,lon1] = flatearthpoly(a1(k),a2(k));
118 [lat2,lon2] = flatearthpoly(b1(h),b2(h));
119 % Use flatearthpoly to convert polygons to Cartesian coordinates
120
121 [latb,lonb] = polybool('intersection',lat1,lon1,lat2,lon2);
122 % detecte intersection entre les enveloppes convexes
123
124
125 Ptint=[intx inty];  %inter
```

```
126
127  Ptint2 = [latb,lonb];  % inclu
128
129  test=(isempty(Ptint)+isempty(Ptint2))/2;
130
131
132  if test==1   %   pas d'intersection ni d'inclusion
133      %disp('vide')
134      bol=1;
135  else
136      %disp('inter')
137  end
138
139
140  end % while
141
142  bol=0;
143
144  plot(lat1,lon1,'b',a1,a2,'b*')
145  hold on
146  plot(lat2,lon2,'r',b1,b2,'r*')
147  hold on
148  patch(latb,lonb,'g')
149  hold on
150  plot(intx,inty,'ko')
151
152  if aire ==1              % Si cas particulier alors une enveloppe convexe = 0
153      aire1=0;
154      [h,aire2] = convhull(b1,b2);
155  elseif aire == 2
156      [k,aire1] = convhull(a1,a2);
157      aire2=0;
158  elseif aire ==3
159      [k,aire1] = convhull(a1,a2);
160      [h,aire2] = convhull(b1,b2);
161  end % if
162
163  statper = ((aire1+aire2)/airetot) ;              %stat
164  fprintf(Doc,'\n');
165  fprintf(Doc,'%2.8f',statper);  %indique le résultat dans "résultat.doc"
166
167  % Test qui fournit le nombre de tper se trouvant dans premier deuxieme troisieme
168  % quatrieme cinquieme
```

```
169  if statper < - abs(statinit)
170      premier = premier + 1;
171  elseif statper == - abs(statinit)
172     deuxieme = deuxieme + 1;
173  elseif statper ==  abs(statinit)
174     quatrieme = quatrieme + 1;
175  elseif statper >  abs(statinit)
176    cinquieme =  cinquieme + 1;
177  else
178      troisieme = troisieme +1;
179  end
180
181  end %for
182
183  disp('la statistique initiale est égale à ')
184  disp(statinit)
185
186  temps=cputime
187
188  disp('sper < sobs')
189  disp(troisieme)
190  disp('sper = sobs')
191  disp(quatrieme )
192  disp('sper > sobs')
193  disp(cinquieme )
194
195
196  fclose(Doc);
```

A.6 chi2test.m (G. Levin, 2003)

```
1  function H=chi2test(x, alpha)
2  %
3  % CHI2TEST: Single sample Pearson Chi Square goodness-of-fit hypothesis test.
4  % H=CHI2TEST(X,ALPHA) performs the particular case of Pearson Chi Square
5  % test to determine whether the null hypothesis of composite normality PDF is
6  % a reasonable assumption regarding the population distribution of a random
7  % sample X with the desired significance level ALPHA.
8  %
9  % H indicates the result of the hypothesis test according to the MATLAB rules
10 % of conditional statements:
11 % H=1 => Do not reject the null hypothesis at significance level ALPHA.
12 % H=0 => Reject the null hypothesis at significance level ALPHA.
13 %
14 % The Chi Square hypotheses and test statistic in this particular case are:
15 %
16 % Null Hypothesis:        X is normal with unknown mean and variance.
17 % Alternative Hypothesis: X is not normal.
18 %
19 % The random sample X is shifted by its estimated mean and normalized by its
20 % estimated standard deviation. The tested bins XP of the assumed normal
21 % distribution are chosen [-inf, -1.6:0.4:1.6, inf] to avoid unsufficient
22 % statistics.
23 %
24 % Let E(x) be the expected frequency X falls within XP according to the normal
25 % distribution and O(x) be the observed frequency. The Pearson statistic,
26 % X2=SUM((E(x)-O(x))^2/E(x)) distributes Chi Square with length(XP)-3 degrees
27 % of freedom.
28 %
29 % The decision to reject the null hypothesis is taken when the P value
30 % (probability that Chi2 random value with length(XP)-3 degrees of freedomd
31 % is greater than X2) is less than % significance level ALPHA.
32 %
33 % X must be a row vector representing a random sample. ALPHA must be a scalar.
34 % The function doesn't check the formats of X and ALPHA, as well as a number of
35 % the input and output parameters.
36 %
37 % The asymptotic limit of the Chi Square test presented is reached when
38 % LENGTH(X)>90.
39 %
```

```
%Acknowledge: Dr. S. Loyka
%
%Author: G. Levin, May, 2003.
%
%References:
% W. T. Eadie, D. Drijard, F. E. James, M Roos and B. Sadoulet,
% "Statistical Methods in Experimental Physics", North-Holland,
% Sec. Reprint, 1982.

%Normalize x
N=length(x);
x=(x-mean(x))/std(x); %standardization

xp=[-inf, -1.6:.4:1.6, inf]; %tested bins
E=0.5*N*diff(erfc(-xp/sqrt(2))); %expected frequency
S=histc(x, xp);
O=S(1:end-1); %%observed frequency
%plot(xp(2:end),E,'k-',xp(2:end),O,'k.');
x2=sum((E-O).^2./E); %statistics

pval=1-gammainc(x2/2,(length(O)-3)/2); %p value

H=(pval>=alpha);
```

Bibliographie

[1] A.Hardy. *Syllabus de cours, Statistiques* , FUNDP, Namur, 2002-2003.

[2] A.Hardy. *Syllabus de cours, Aspects statistiques de la Classification*, FUNDP, Namur, 2005-2006.

[3] A.Hardy. *Validation of a clusturing structure : determination of the number of clusters*, Wriley, 2007 (à paraître).

[4] A.Hardy, J.P.Rasson. Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des Données*, 1982, pages.41-56.

[5] A.Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, 1996, pages.83-96.

[6] J.P.Rasson, V.Granville. .Geometrical tools in classification. *Computational Statistics and Data Analysis*, 1996, pages.105-123.

[7] B.D.Ripley, J.P.Rasson. Finding the edge of a Poisson forest. *Journal of Applied Probability*, 1977, pages.483-491.

[8] M.Moore. On the estimation of a convex set. *Annals of Statistics*, 1984, pages 1090-1099.

[9] Cox, Isham. *Point processes*, Chapman and Hall, 1980.

[10] P.Legendre and L.Legendre. *Numerical Ecology*, Elsevier, second english edition, 1998.

[11] P.Legendre. *Biostatistique I, chapitre 6 : L'inférence statistique : les tests d'hypothèse*,
http ://www.bio.umontreal.ca/legendre/BIO2041/index.html (consulté le 23 septembre 2005)

[12] B.Manly. *Randomization bootstrap and Monte Carlo methods in biology*, Chapman and Hall, second edition, 1997.

[13] R.A.Fisher. *Statistical methods for research workers*, Oliver and Boyd, 1925.

[14] R.A.Fisher. *The design of experiments*, Oliver and Boyd, 1935.

[15] E.L.Lehmann. *Testing Statistical Hypothesis*, second edition. New York, 1986

[16] Agro-Montpellier, Tables Statistiques, http ://www.agro-montpellier.fr/cnam-lr/statnet/tables.htm

[17] John C. Pezzullo, Probability Distribution Functions, http ://statpages.org/pdfs.html

[18] SAS. *User's Guide*, SAS institute Inc, NC, USA, 1990.

[19] Matlab 7.0.0. Help Navigator, 2004